UNIVERSIDADE DE SÃO PAULO
ESCOLA DE ARTES, CIÊNCIAS E HUMANIDADES
PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO

RAFAEL LUIZ TESTA

Síntese de expressões faciais em fotografias para representação de emoções

São Paulo

2019

RAFAEL LUIZ TESTA

Síntese de expressões faciais em fotografias para representação de emoções

Dissertação apresentada à Escola de Artes, Ciências e Humanidades da Universidade de São Paulo para obtenção do título de Mestre em Ciências pelo Programa de Pós-graduação em Sistemas de Informação.

Área de concentração: Metodologia e Técnicas da Computação

Versão corrigida contendo as alterações solicitadas pela comissão julgadora em 04 de dezembro de 2018. A versão original encontra-se em acervo reservado na Biblioteca da EACH-USP e na Biblioteca Digital de Teses e Dissertações da USP (BDTD), de acordo com a Resolução CoPGr 6018, de 13 de outubro de 2011.

Orientador: Profa. Dra. Fátima de Lourdes dos Santos Nunes Marques

Coorientador: Profa. Dra. Arianne Machado Lima

São Paulo

2019

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

CATALOGAÇÃO-NA-PUBLICAÇÃO

(Universidade de São Paulo. Escola de Artes, Ciências e Humanidades. Biblioteca)
CRB 08-4936

Testa, Rafael Luiz

Síntese de expressões faciais em fotografias para representação de emoções / Rafael Luiz Testa ; orientadora, Fátima de Lourdes dos Santos Nunes Marques ; coorientadora, Arianne Machado Lima. – 2019.
104 p. : il.

Dissertação (Mestrado em Ciências) - Programa de Pós-Graduação em Sistemas de Informação, Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, em 2018
Versão corrigida

1. Reconhecimento de padrões. 2. Reconhecimento de imagem. 3. Emoções. 4. Percepção de face. I. Marques, Fátima de Lourdes dos Santos Nunes, orient. II. Machado-Lima, Arianne, coorient. III. Título

CDD 22.ed. – 006.42

Agradecimentos

Várias pessoas colaboraram de diferentes formas, o que ajudou a facilitar a realização desta pesquisa e, por isso, sou grato. Aos Professores Anastasios Delopoulos, Jesús P. Mena-Chalco e Luiz Velho pelas imagens disponibilizadas. Aos meus pais pelo contínuo apoio às minhas escolhas profissionais. Aos meus colegas de laboratório Leonardo, Renan, Leila e Mateus pelas inúmeras contribuições no dia a dia do trabalho. Às minhas orientadoras Fátima e Ariane por todo esforço e dedicação que contribuíram para um trabalho melhor. E um agradecimento especial a Tati, pois, ela não só ajudou a tornar este trabalho melhor, como também me ajudou a tornar-me uma pessoa melhor. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

“Os movimentos expressivos conferem vividez e energia às palavras ... Revelam pensamentos e intenções com maior veracidade as palavras, que ser objeto de simulação”
(Darwin, 1872/1965, p. 364)

Resumo

TESTA, Rafael Luiz. **Síntese de expressões faciais em fotografias para representação de emoções**. 2019. 104 f. Dissertação (Mestrado em Ciências) – Escola de Artes, Ciências e Humanidades, Universidade de São Paulo, São Paulo, 2018.

O processamento e a identificação de emoções faciais constituem ações essenciais para estabelecer interação entre pessoas. Alguns transtornos psiquiátricos podem limitar a capacidade de um indivíduo em reconhecer emoções em expressões faciais. De modo a contribuir com a solução deste problema, técnicas computacionais podem ser utilizadas para compor ferramentas destinadas ao diagnóstico, avaliação e treinamento no reconhecimento de tais expressões. Com esta motivação, o objetivo deste trabalho é definir, implementar e avaliar um método para sintetizar expressões faciais que representam emoções em imagens de pessoas reais. Nos trabalhos encontrados na literatura a principal ideia é que a expressão facial da imagem de uma pessoa pode ser reconstituída na imagem de outra pessoa. Este estudo difere-se das abordagens apresentadas na literatura ao propor uma técnica que considera a similaridade entre imagens faciais para escolher aquela que será empregada como origem para a reconstituição. Desta maneira, pretende-se aumentar o realismo das imagens sintetizadas. A abordagem sugerida para resolver o problema, além de buscar as faces mais similares em banco de imagens, faz a deformação dos componentes faciais e o mapeamento das diferenças de iluminação na imagem destino. O realismo das imagens geradas foi mensurado de forma objetiva e subjetiva usando imagens disponíveis em bancos de imagens públicos. Uma análise visual mostrou que as imagens sintetizadas com base em faces similares apresentaram um grau de realismo adequado, principalmente quando comparadas com imagens sintetizadas a partir de faces aleatórias. Além de constituir uma contribuição para a geração de imagens a serem aplicadas em ferramentas de auxílio ao diagnóstico e terapia de distúrbios psiquiátricos, oferece uma contribuição para a área de Ciência da Computação, por meio da proposição de novas técnicas de síntese de expressões faciais.

Palavras-chaves: Síntese de Expressão Facial. Mapeamento de Expressão Facial. Clonagem de Expressão Facial. Transferência de expressão. Geração de expressão facial.

Abstract

TESTA, Rafael Luiz. **Facial expression synthesis in photographs for emotion representation**. 2019. 104 p. Dissertation (Master of Science) – School of Arts, Sciences and Humanities, University of São Paulo, São Paulo, 2018.

The ability to process and identify facial emotions are essential factors for an individual's social interaction. Some psychiatric disorders can limit an individual's ability to recognize emotions in facial expressions. This problem could be confronted by using computational techniques in order to develop learning environments for diagnosis, evaluation, and training in identifying facial emotions. With this motivation, the objective of this work is to define, implement and evaluate a method to synthesize realistic facial expression that represents emotions in images of real people. The main idea of the studies found in the literature is that a facial expression of one person's image can be reenacted in an another person's image. The study differs from the approaches presented in the literature when proposing a technique that considers the similarity between facial images to choose the one that will be used as the origin for reenactment. As a result, we intend to increase the realism of the synthesized images. Our approach to solve the problem, besides searching for the most similar facial components in the image dataset, also deforms the facial elements and maps the differences of illumination in the target image. A visual analysis showed that the images synthesized on the basis of similar faces presented an adequate degree of realism, especially when compared with images synthesized from random faces. The study will contribute to the generation of the images applied to tools for the diagnosis and therapy of psychiatric disorders, and also contribute to the computational field, through the proposition of new techniques for facial expression synthesis.

Keywords: Facial Expression Synthesis Facial Expression Mapping. Facial Expression Cloning. Expression Transfer. Facial Expression Generation.

Lista de figuras

Figura 1 – Origens das expressões faciais.	23
Figura 2 – Interação entre as grandes áreas da computação gráfica e a sua relação com dados e imagens.	24
Figura 3 – Relação do trabalho com as grandes áreas da computação gráfica. A área destacada em vermelho mostra as áreas em que o projeto está inserido.	24
Figura 4 – <i>Continuum</i> das abordagens baseadas em imagem de acordo com a geometria.	25
Figura 5 – Exemplo de subjanelas positivas para detecção da face.	25
Figura 6 – Exemplos de subjanelas e seus f -scores. Um método guloso retornaria uma solução subótima com a janela do centro (f score=7), enquanto a solução ideal para essas três janelas seria as duas do canto (f score=12).	26
Figura 7 – Exemplo de landmarks faciais para os olhos e a sobrancelha.	27
Figura 8 – Exemplo de triangulação de <i>Delaunay</i>	29
Figura 9 – Exemplo de transformação afim.	29
Figura 10 – Entrada para as redes neurais tradicionais e convolutivas.	31
Figura 11 – Ilustração da arquitetura de uma <i>CNN</i>	32
Figura 12 – Bloco residual em uma ResNet.	32
Figura 13 – Gráfico da porcentagem média do atendimento aos critérios de qualidade.	37
Figura 14 – Diagrama de <i>Venn</i> da distribuição das abordagens.	37
Figura 15 – Estágios da síntese de expressões faciais por abordagem. Linhas tracejadas indicam um estágio opcional. Retângulos com linhas tracejadas indicam estágios opcionais e setas com linhas tracejadas indicam que apenas um dos caminhos deve ser seguido.	38
Figura 16 – Relação entre as diferentes fases do projeto.	47
Figura 17 – Exemplo de questão exibida aos participantes.	54
Figura 18 – Diagrama das etapas de sintetização pela presente proposta.	57

Figura 19 – Ilustração das arquiteturas das redes <i>CNN</i> para detecção facial e <i>ResNet</i> para descrição da face. “Conv” significa convolução, “TC” representa as camadas totalmente conectadas, o número anterior à “conv” representa o tamanho da janela do filtro, o posterior à “conv” quantidade de filtros, “/2” trata-se do <i>pooling</i> e as linhas tracejadas mostram os atalhos com aumento de dimensão.	59
Figura 20 – Identificação do <i>landmarks</i> faciais com exemplos. Os pontos em azul são os <i>landmarks</i> identificados, enquanto os pontos em vermelho foram adicionados com base na posição dos demais pontos.	60
Figura 21 – Triangulação da face por <i>Delaunay</i> e ajustes manuais. Os triângulos destacados em vermelho na imagem da esquerda foram ajustados manualmente e conforme os triângulos em vermelho na direita.	61
Figura 22 – Esquematização das principais etapas do alinhamento para uma face. .	62
Figura 23 – Esquematização das etapas empregadas para encontrar uma imagem facial similar.	65
Figura 24 – Esquematização da deformação da face neutra para expressiva. Ilustrações e exemplos esquematizam as etapas de cálculo dos <i>landmarks</i> expressivos, deformação da face, deformação da parte oclusa e mesclagem da parte oclusa.	67
Figura 25 – Esquematização das etapas empregadas para mapeamento das características, envolvendo a imagem razão, melhorias nas imagem razão e a síntese da imagem expressiva. A imagem razão foi invertida para uma melhor visualização, ou seja, os <i>pixels</i> pretos equivalem a um e os <i>pixels</i> brancos a zero.	69
Figura 26 – Porcentagem média de atribuições como “real” e sintética (“não real”) para cada uma das imagens.	72
Figura 27 – Exemplos de imagens com e sem bordas irregulares na região da boca e em regiões com descontinuidade.	75
Figura 28 – Exemplos de imagens com diferença de cor. Imagens retiradas ou sintetizadas a partir de imagens de IMPA Faces.	76
Figura 29 – Exemplos de componentes faciais com tamanhos não naturais.	77
Figura 30 – Exemplos de expressões faciais com e sem enrugamentos.	77

Figura 31 – Porcentagem de respostas como “real” e sintéticas (“não real”) para cada expressão facial.	78
Figura 32 – Exemplos de expressões faciais.	78
Figura 33 – Exemplos de imagens anteriores e posteriores as melhorias em relação às bordas irregulares proeminentes.	79
Figura 34 – Exemplos de imagens anteriores e posteriores a adição dos enrugamentos nas imagens sintéticas.	80
Figura 35 – Exemplos de imagens anteriores e posteriores a triangulação manual.	82
Figura 36 – Exemplos de imagens anteriores e posteriores ao mascaramento. Os enrugamentos das figuras a e b foram sintetizados a partir da face deformada apresentada na imagem da figura 34b.	83
Figura 37 – Exemplo de imagens sintéticas de satisfação para imagens selecionadas pelas abordagens características faciais, histograma de cor, e aleatória. À direita são apresentadas a face destino neutra e <i>ground truth</i>	84
Figura 38 – Exemplo de imagens sintéticas de surpresa para imagens selecionadas pelas abordagens características faciais, histograma de cor, e aleatória. À direita são apresentadas a face destino neutra e <i>ground truth</i>	85
Figura 39 – Exemplo de detecção de erro na posição dos <i>landmarks</i> faciais para região da sobrancelha.	85
Figura 40 – Exemplos de faces sintéticas para as três diferentes abordagens com variados indivíduos representando as seis expressões faciais consideradas universais.	87
Figura 41 – Gráfico exemplificando a precisão e revocação para quatro imagens de um indivíduo com o descritor de histograma de cor.	89
Figura 42 – Gráfico da precisão média por imagem encontrada para cada descritor.	90
Figura 43 – Exemplo das distâncias utilizadas para compor a métrica de avaliação dos <i>landmarks</i> faciais.	91
Figura 44 – Imagem diferença e imagem resultante do SSIM para um mesmo indivíduo considerando a diferença entre as imagens sintéticas (seleção das imagens fontes de modo aleatório, por histograma de cor e características faciais) e a imagem <i>ground truth</i> . A coluna <i>ground truth</i> mostra o resultado esperado para cada uma das imagens da linha correspondente.	93

Lista de tabelas

Tabela 1 – Critérios de inclusão e exclusão	35
Tabela 2 – Critérios de qualidade separados por fator e porcentagem de atendimento aos mesmos.	36
Tabela 3 – Características dos bancos de imagens utilizados nos experimentos. . .	51
Tabela 4 – Matriz de confusão das respostas. As linhas referem-se ao tipo de imagem (sintéticas ou reais) e a todas as imagens independentemente do tipo, enquanto as colunas referem-se às respostas (sintéticas ou reais) e ao total de respostas.	73
Tabela 5 – Comparação das porcentagens de atribuição de realismo das quatro imagens com melhores resultados obtidos (ver figura 26) em relação à avaliação dos quatro vídeos de Averbuch-Elor et al. (2017).	74
Tabela 6 – Comparativo das áreas sob a curva da <i>ground truth</i> e imagem sintética para a seleção da imagens utilizando histograma de cor, características faciais e de forma aleatória nos bancos de expressão facial IMPA Faces e MUG.	92
Tabela 7 – Comparativo das áreas sob a curva da imagem sintética e face destino neutra para a seleção da imagens utilizando histograma de cor, caracte- rísticas faciais e de forma aleatória nos bancos de expressão facial IMPA Faces e MUG.	94

Lista de abreviaturas e siglas

3D	Tridimensional
CK+	<i>Extended Cohn-Kanade Dataset</i>
CNN	<i>Convolutional Neural Network</i>
DP	<i>Desvio Padrão</i>
ERI	<i>Expression ratio image</i>
FFPD	<i>Facial Feature Point Detection</i>
F_{fn}	Face Fonte Neutra
F_{fe}	Face Fonte Expressiva
F_{dn}	Face Destino Neutra
F_s	Face Sintética
FSeN	Face Seleccionada Neutra
FSeE	Face Seleccionada Expressiva
FSiE	Face Sintetizada Expressiva
HHD	<i>Helmholtz-Hodge decomposition</i>
IMPA FACES	<i>Instituto de Matemática Pura e Aplicada facial expression database</i>
KNN	k vizinhos mais próximos
JAFFE	<i>Japanese Female Facial Expression database</i>
LFW	<i>Labeled Faces in the Wild face database</i>
ldif	Diferenças entre <i>landmarks</i>
MSE	<i>Mean Squared Error</i>
MUG Database	<i>Multimedia Understanding Group facial expression Database</i>
PCA	<i>Principal Component Analysis</i>

PSNR	<i>Average Peak Signal to Noise Ratio</i>
ResNet	<i>Residual Neural Network</i>
RS	Revisão Sistemática
SLPP	<i>Supervised Locality Preserving Projections</i>
SSIM	<i>Structural SIMilarity</i>
SVD	<i>Singular Value Decomposition</i>
TPS	<i>Thin Plate Spline</i>
VGG Face Dataset	Visual Geometry Group Face Dataset
YCbCr	Luminância; Croma: Azul; Croma: Vermelha
Y_{dd}	Luminância da face destino deformada
Y_{dn}	Luminância da face destino neutra
Y_{fn}	Luminância da face fonte neutra
Y_{fe}	Luminância da face fonte expressiva
Y_s	Luminância da face sintética

Sumário

1	Introdução	16
1.1	<i>Motivação</i>	17
1.2	<i>Justificativa</i>	18
1.3	<i>Objetivos e hipótese</i>	19
1.4	<i>Resumo do método desenvolvido e contribuições</i>	20
1.5	<i>Organização do documento</i>	21
2	Conceitos fundamentais	22
2.1	<i>Expressões faciais</i>	22
2.2	<i>Sintetização de imagens</i>	22
2.2.1	<i>Síntese baseada em imagens</i>	23
2.3	<i>Detecção de Objetos por Margem Máxima</i>	25
2.4	<i>Landmarks</i>	26
2.5	<i>Função de similaridade</i>	27
2.6	<i>Triangulação</i>	28
2.7	<i>Transformações afins</i>	28
2.8	<i>Adequação da iluminação</i>	29
2.9	<i>Redes neurais convolutivas</i>	30
2.9.1	<i>Redes Neurais Residuais</i>	32
2.10	<i>Métricas para comparação de imagens</i>	33
2.11	<i>Considerações finais</i>	34
3	Levantamento bibliográfico	35
3.1	<i>Pré-processamento</i>	37
3.2	<i>Síntese baseada em Deformação e Mapeamento</i>	39
3.3	<i>Síntese baseada em Aprendizado de máquina</i>	40
3.4	<i>Sintetização com outras abordagens</i>	41
3.5	<i>Avaliação dos resultados</i>	41
3.5.1	<i>Avaliação dos resultados pelos autores</i>	42
3.6	<i>Bancos de imagens de expressões faciais</i>	42
3.7	<i>Discussões</i>	43

3.7.1	Métodos de pré-processamento	43
3.7.2	Métodos de síntese	44
3.7.3	Métricas de avaliação dos resultados	45
3.7.4	Bancos de imagens de expressões faciais	46
3.8	<i>Conclusões</i>	46
4	Materiais e métodos	47
4.1	<i>Desenvolvimento do método proposto</i>	47
4.1.1	Seleção das imagens fonte	48
4.1.2	Deteção facial	49
4.1.3	Correspondência entre características faciais	49
4.1.4	Alinhamento das imagens faciais	49
4.1.5	Deformação e Mapeamento	50
4.2	<i>Avaliação dos resultados</i>	50
4.2.1	Bancos de expressões faciais	50
4.2.2	Métricas objetivas	52
4.2.3	Métrica subjetiva	54
4.3	<i>Considerações finais</i>	56
5	Síntese de expressões faciais por faces similares	57
5.1	<i>Deteção facial</i>	57
5.2	<i>Deteção dos landmarks faciais</i>	58
5.3	<i>Triangulação das faces</i>	60
5.4	<i>Alinhamento das faces</i>	61
5.5	<i>Seleção de imagens fonte</i>	63
5.6	<i>Deformação da face</i>	66
5.7	<i>Mapeamento da iluminação</i>	68
5.8	<i>Considerações finais</i>	70
6	Resultados e discussões	71
6.1	<i>Avaliação da percepção de realismo</i>	71
6.1.1	Análise estatística	71
6.1.2	Fatores que influenciam a percepção do realismo	74
6.1.3	Análise das expressões faciais	77

6.1.4	Melhorias após experimento	79
6.1.5	Demais melhorias	81
6.2	<i>Imagens sintetizadas</i>	82
6.3	<i>Experimento para a avaliação da busca por face similar</i>	88
6.4	<i>Experimento para avaliação da síntese com face similar</i>	89
6.5	<i>Experimento para avaliação da preservação da identidade</i>	92
7	Conclusões e trabalhos futuros	95
	Referências¹	97

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

1 Introdução

A comunicação de uma emoção pode ser estabelecida por meio verbal (SCHERER, 2003) e também por meios não verbais, tais como expressões faciais, posturas e gestos. A comunicação efetiva de uma emoção é benéfica tanto para quem está se comunicando quanto para quem está tentando compreender (HESS, 2001).

A capacidade de processar e identificar emoções faciais é um componente essencial da comunicação humana e da interação social de acordo com as normas e costumes culturais, baseadas em padrões comuns a cada cultura. Embora não seja unânime, boa parte dos pesquisadores da área consideram seis emoções universais já estabelecidas: satisfação, tristeza, raiva, aversão, medo e surpresa (EKMAN; FRIESEN, 1971). Cada uma dessas emoções corresponde a um padrão específico de movimento dos músculos da face e envolve neurocircuitos parcialmente separáveis, capazes de preparar uma resposta comportamental apropriada (ROCCA et al., 2009; GOSSELIN; KIROUAC et al., 1995).

Segundo a revisão de Rocca et al. (2009), o processamento de expressões faciais das emoções, que compreende a interpretação do estímulo para a preparação de uma resposta comportamental apropriada, é apoiado por sistemas neurais específicos. Estudos de imagens cerebrais sugerem que regiões específicas no córtex temporal inferior medeiam a percepção de faces. Doenças distintas podem se caracterizar por déficits diferentes em aspectos específicos do processamento emocional. Por exemplo, pacientes com doença de Urbach-Wiethe, que causa dano bilateral na amígdala, são incapazes de distinguir medo tanto em faces reais quanto em imaginárias. No estudo dos transtornos psiquiátricos, o paradigma de reconhecimento de faces vem recebendo maior atenção nos últimos anos, especialmente na investigação de transtornos do espectro autista (HARMS; MARTIN; WALLACE, 2010), transtornos do humor (ROCCA et al., 2009) e esquizofrenia (TAYLOR; MACDONALD, 2012). Estudos mais recentes têm avaliado a possibilidade de reconhecimento de padrões em populações de risco para transtornos mentais (MOURÃO-MIRANDA et al., 2012). Portanto, o reconhecimento de expressões faciais tem sido empregado como um instrumento na avaliação de transtornos mentais e também como uma ferramenta para treinamento dessa habilidade.

Estudos sugerem que indivíduos portadores de transtornos do espectro autista apresentam dificuldades na interação e comunicação social (NEWSCHAFER; CURRAN,

2003), incluindo a habilidade de identificar a expressão facial das emoções em outras pessoas (CLARK; WINKIELMAN; MCINTOSH, 2008). Neste contexto, imagens estáticas de expressões faciais sendo encenadas por atores foram incluídas em jogos interativos para o treinamento dessas habilidades, e os pacientes mostraram melhorias significativas em relação a um grupo controle (GOLAN; BARON-COHEN, 2006). Além disso, o uso de ambientes de realidade virtual, por meio de modelos tridimensionais (3D) para representação de expressões faciais em ambiente multimídia interativo e de *software* de treinamento utilizando interface humano-computador apresentaram resultados positivos no que se refere à melhora desses aspectos em indivíduos do espectro autista (LAHIRI et al., 2013; CHENG; LING, 2008; GRZYNSZPAN; MARTIN; NADEL, 2008). Assim, aplicações de processamento gráfico podem contribuir no treinamento do reconhecimento de expressões faciais de emoções.

1.1 Motivação

Em geral, as ferramentas para treinamento da habilidade de reconhecimento de emoções apresentam imagens provenientes de bases de atores simulando diferentes emoções. Além da dificuldade de simular algumas emoções realisticamente (EKMAN; HAGER; FRIESEN, 1981), esta abordagem pode apresentar limitações ao diagnóstico e ao treinamento, pois, estudos têm mostrado, por exemplo, que pessoas com transtorno do espectro autista têm mais dificuldade em identificar emoções em faces não familiares (PIERCE et al., 2004). Portanto, faces familiares podem ser utilizadas para a composição de diferentes níveis de dificuldade em um sistema de treinamento, o que pode contribuir para a flexibilização dessas ferramentas.

Além disso, as dificuldades de reconhecimento variam de acordo com cada paciente, podendo, por exemplo, um determinado paciente não identificar expressões faciais específicas ou variações mais sutis. Uma abordagem alternativa seria possibilitar ao terapeuta a geração de imagens com expressões faciais que representam diferentes intensidades de variadas emoções, de acordo com a necessidade de cada paciente. Dessa forma, a geração de imagens faciais para representar emoções pode contribuir para que ferramentas de auxílio ao diagnóstico e treinamento de pacientes sejam adaptadas às necessidades de cada indivíduo.

Com esta motivação, o presente trabalho faz parte de um projeto de pesquisa mais abrangente que visa a desenvolver jogos para reconhecimento de expressões faciais que possam ser explorados no campo de pesquisa dos transtornos mentais. Além da necessidade de gerar imagens com diferentes intensidades de emoções, também é de interesse que as emoções possam ser sintetizadas em imagens pré-selecionadas, o que viabiliza, por exemplo, considerar faces familiares aos pacientes. Sendo assim, grande parte do desafio computacional da implementação de tais jogos reside na geração destas imagens.

Além do contexto de jogo, a síntese de expressões faciais tem aplicações em automação de agentes web interativos para videoconferências com baixa largura de banda (LI; CHANG; CHANG, 2007), interfaces interativas (MENDI; BAYRAK, 2011), melhoria em fotografias (FUJISHIRO et al., 2009), animação computacional (NOH; NEUMANN, 2001) e ícones personalizados (LI; CHANG; CHANG, 2007).

As abordagens de síntese de expressões faciais também têm sido empregadas em aplicações mais complexas, como o planejamento de cirurgia facial (KEEVE et al., 1998) e o aumento de dados de treinamento (MOHAMMADIAN; AGHAEINIA; TOWHIDKHAH, 2016). Adicionalmente, têm sido inspiração para abordagens de áreas correlacionadas, como o desenvolvimento de novas técnicas de reconhecimento de emoções em expressões faciais (SONG et al., 2010), constituir parte de um sistema de reconhecimento facial invariante a expressão facial (AMBERG; KNOTHE; VETTER, 2008) ou compor um sistema de envelhecimento/rejuvenescimento de idade que considera a expressão facial (ZHANG; SONG; QI, 2017).

1.2 Justificativa

O processo mais empregado para a síntese de expressões faciais consiste nas seguintes etapas: i) identificação e extração das características expressivas em um par de imagens fontes: uma contendo a expressão facial desejada (Face Fonte Expressiva - F_{fe}) e outra face do mesmo indivíduo sem a expressão (Face Fonte Neutra - F_{fn}); ii) adaptação e mesclagem das características identificadas a uma imagem destino sem a expressão (Face Destino Neutra - F_{dn}), isto é, a imagem da face que se deseja reconstituir a expressão (ZHANG et al., 2014). Esse processo altera a F_{dn} para uma configuração expressiva, gerando uma

imagem sintética (Face Sintética - F_s) que possui a identidade de F_{dn} e a expressão de F_{fe} . Uma descrição mais detalhada desse processo é apresentada nos capítulos 3 e 4.

As abordagens encontradas na literatura podem ser divididas, principalmente, em dois grupos: “Deformação e Mapeamento” (Seção 3.2) e “Aprendizado de Máquina” (Seção 3.3). Ambos os grupos possuem algumas limitações.

O primeiro grupo de abordagens usa técnicas de deformação e mapeamento dos componentes faciais, obtendo tais componentes na imagem origem (F_{fe}) e adequando-os à imagem destino (F_{dn}). Esta abordagem é bastante dependente da imagem fonte (Seção 3.8) e a escolha de uma imagem inadequada pode gerar inconsistências na imagem sintetizada (XIONG et al., 2010). Poucos trabalhos se preocuparam em selecionar essas imagens de forma a otimizar seus resultados, como apresentado na seção 3.1.

No segundo grupo de abordagens, técnicas de aprendizado de máquina usam todas as imagens para construir seu modelo e sintetizar a nova expressão facial. Neste caso, é difícil sintetizar expressões faciais quando há poucos exemplos para o treinamento (Seção 3.7), o que pode ser um problema em expressões faciais menos comuns (XIONG et al., 2010) e espontâneas (ZHOU; SHI, 2017).

A abordagem adotada no presente projeto tem como foco a reprodução realista de expressões faciais em fotografias, usando, para isso, um processo de seleção de imagens fonte que visa a encontrar, no banco de imagens, a F_{fn} mais similar à F_{dn} .

1.3 Objetivos e hipótese

Com a finalidade de minimizar as limitações apresentadas, a presente proposta pretende aprimorar a seleção das imagens fonte (Seção 3.1), possibilitando selecionar uma face fonte que melhor se adeque à face destino, independentemente da quantidade de exemplos disponíveis. Para isso, foi considerada uma base de imagens previamente construída contendo expressões faciais das diferentes emoções, assim como a expressão neutra da mesma pessoa.

Assim, este projeto de pesquisa tem como objetivo o desenvolvimento, a implementação e a avaliação de uma abordagem para síntese realista de expressões faciais representando emoções, considerando especificamente a seleção eficiente de faces fontes similares. No contexto do presente projeto, o realismo está relacionado com o fato de

obter uma imagem sintetizada de tal forma que os componentes inseridos ou alterados (por exemplo, boca, olhos e rugas) mantenham as mesmas características de uma imagem facial expressiva real.

Para alcançar esse objetivo geral são estabelecidos os seguintes objetivos específicos:

- definir e implementar um método que estabeleça uma correspondência, entre regiões faciais em duas faces diferentes;
- definir e implementar um método que permita encontrar no banco de imagens a face de expressão neutra mais similar à imagem destino;
- definir e implementar um método para alterar ou acrescentar componentes faciais na imagem destino, visando a obter imagens finais realistas;
- definir e implementar um método capaz de produzir detalhes finos nas imagens, como enrugamentos;
- validar os métodos desenvolvidos de forma objetiva;
- avaliar o realismo por meio de experimento com usuários.

1.4 *Resumo do método desenvolvido e contribuições*

O método desenvolvido possui seis etapas, sendo elas: reconhecimento facial, correspondência entre as características faciais, alinhamento facial, seleção da imagem fonte, deformação dos componentes faciais e mapeamento da iluminação. O reconhecimento facial é realizado com uma *CNN* (*Convolutional Neural Network*) pré-treinada. Os *landmarks* faciais são detectados com um modelo de árvores de regressão cascata pré-treinado, para estabelecer a correspondência entre as características faciais de duas faces diferentes. O alinhamento facial rotaciona, escala e corta a imagem com base no ângulo e na distância entre os olhos. A seleção das imagens fontes compara as características faciais extraídas da imagem por uma *CNN* pré-treinada para encontrar a imagem mais similar. A deformação dos componentes faciais estima uma forma expressiva para a face destino e deforma os componentes faciais por meio de transformações afins. Por fim, a razão entre as imagens fontes é mapeada na face destino para transpor detalhes mais finos relacionados à iluminação.

A abordagem resultante obteve um grau de realismo adequado, avaliado por meio de um experimento com voluntários e uso de métricas objetivas citadas na literatura. A principal contribuição do trabalho é a adição da etapa de seleção automática da imagem

fonte que considera a similaridade entre as faces; em consequência dessa contribuição, define-se que há uma segunda importante contribuição em relação ao tamanho do conjunto de imagens fontes, pois não é necessário um conjunto de imagens fontes com muitos exemplares e nem a obtenção de imagens da pessoa destino. O trabalho também resultou nas seguintes contribuições secundárias:

- melhorias no processo de deformação e mapeamento;
- definição de um método de avaliação objetiva a partir de métricas utilizadas na literatura;
- definição de um método de avaliação subjetiva para avaliar o realismo de imagens faciais sintéticas.

A partir das considerações tecidas, a hipótese a ser verificada por este projeto é: “a seleção de uma face fonte mais similar à face destino resulta em imagens sintetizadas expressivas com nível satisfatório de realismo”.

1.5 Organização do documento

Além deste capítulo introdutório, este documento é composto pelos seguintes capítulos: no capítulo 2 é apresentado um resumo dos conceitos básicos sobre síntese de imagens e expressões faciais; o capítulo 3 mostra brevemente um levantamento bibliográfico resultante de uma revisão sistemática; o capítulo 4 apresenta a abordagem proposta e a forma de avaliação dos resultados; o capítulo 6 expõe e discute os resultados referentes às avaliações por métricas objetivas e uma subjetiva, além disso, exemplos de imagens sintetizadas são utilizados para discutir os resultados obtidos; por fim, o capítulo 7 apresenta as contribuições e trabalhos futuros.

2 Conceitos fundamentais

Dentro do contexto do presente projeto, a síntese das expressões faciais envolve conceitos disponibilizados neste capítulo relacionados às expressões faciais (Seção 2.1), sintetização de imagens (Seção 2.2), identificação dos *landmarks* faciais (Seção 2.4), medidas de similaridade (Seção 2.5), deformação nas imagens (Seção 2.7), alterações de iluminação (Seção 2.8), detecção de faces (Seção 2.9) e extração de características faciais (Seção 2.9). Por fim, na seção 2.10 são descritas métricas para comparação de imagens.

2.1 Expressões faciais

As expressões faciais podem ser entendidas como o conjunto de movimentos faciais e deformações dos componentes faciais (FASEL; LUETTIN, 2003). No contexto da computação gráfica, a síntese de expressões pode ser alcançada com tais modificações nas imagens. Essas modificações devem ser consistentes com as alterações referentes a uma determinada expressão facial, para a reconstituição da expressão facial desejada.

As expressões faciais possuem diversas origens de acordo com os estímulos que as desencadearam. Tais estímulos podem ser estados mentais, elementos da comunicação não verbal, atividades fisiológicas e enfatizadores de aspectos na comunicação não verbal (FASEL; LUETTIN, 2003), conforme ilustrado na figura 1. O presente estudo tem como foco a síntese das expressões faciais ligadas a emoções; entretanto, a abordagem pode ser estendida para a síntese das demais origens das faces expressivas (ver figura 1).

2.2 Sintetização de imagens

A computação gráfica pode ser dividida em três grandes áreas: síntese de imagens, processamento de imagens e análise de imagens (FOLEY, 1996). A síntese de imagens consiste na geração de imagens por computador com base nas especificações geométricas e visuais dos objetos e cenas sintetizados (AZEVEDO; CONCI, 2003). Já o processamento de imagens tem como foco o conjunto de técnicas usadas para capturar, representar e transformar imagens digitais (PEDRINI; SCHWARTZ, 2008). Por fim, a análise de imagens procura extrair das imagens desejadas as características (AZEVEDO; CONCI,

Figura 1 – Origens das expressões faciais.

Fonte: Traduzido de Fasel e Luetlin (2003).

2003), considerando principalmente a textura, a forma, níveis de cinza ou as cores de uma determinada imagem digital (PEDRINI; SCHWARTZ, 2008). A figura 2 ilustra os conceitos descritos ao apresentar as associações entre grandes áreas da computação gráfica e os dados/imagens.

2.2.1 Síntese baseada em imagens

A definição supracitada de síntese de imagens estabelece esse processo a partir da transformação dos dados em imagem (ver figura 2). Entretanto, outra maneira de conseguir as informações sobre os objetos e as cenas é a síntese baseada em imagens (*image-based rendering*), na qual as especificações da cena são obtidas de imagens ou de cenas sintetizadas anteriormente (FOLEY et al., 1990). Portanto, tal abordagem envolve a obtenção dos dados da cena por meio da análise de imagens e a reconstrução por meio da síntese de imagens (MAGNOR, 2005). Além disso, no contexto deste trabalho, a síntese também envolve a área de processamento de imagens porque a obtenção das novas fotografias é realizada por meio da transformação dessas imagens. Portanto, a abordagem

Figura 2 – Interação entre as grandes áreas da computação gráfica e a sua relação com dados e imagens.

Fonte: Adaptado de Conci, Azevedo e Leta (2008).

apresentada no capítulo 4 se insere no contexto das três grandes áreas, conforme ilustrado na figura 3.

Figura 3 – Relação do trabalho com as grandes áreas da computação gráfica. A área destacada em vermelho mostra as áreas em que o projeto está inserido.

Fonte: Adaptado de Conci, Azevedo e Leta (2008).

As abordagens baseadas em exemplos são comumente divididas de acordo com o uso da geometria para sua composição (SHUM; CHAN; KANG, 2008): síntese sem geometria, síntese com geometria implícita e síntese com geometria explícita, conforme ilustrado na figura 4. Enquanto na síntese com geometria explícita as cenas são construídas a partir dos modelos geométricos, na síntese sem geometria é realizada com técnicas como interpolações, *blending*, entre outras (SHUM; CHAN; KANG, 2008). A abordagem intermediária é a síntese com geometria implícita, na qual é necessária uma correspondência entre as imagens

(SHUM; CHAN; KANG, 2008). Esta correspondência pode ser definida, por exemplo, por meio da identificação dos *landmarks* faciais, conforme exposto na seção 5.2.

Figura 4 – *Continuum* das abordagens baseadas em imagem de acordo com a geometria.

Fonte: Adaptado de Shum, Chan e Kang (2008).

2.3 Detecção de Objetos por Margem Máxima

A maior parte dos métodos de detecção de objetos extrai um conjunto de janelas candidatas sobrepostas agrupadas em dois subconjuntos: negativo (sem objeto) e positivo (com objeto) (SIDDIQUI et al., 2016). A figura 5 mostra um conjunto de subjanelas positivas para detecção facial. Idealmente a detecção de objetos deveria operar em todas as possíveis subjanelas para verificar a localização exata de um objeto. Entretanto, operar sobre todo esse conjunto pode ser bastante custoso em tempo computacional. Assim, por questão de eficiência, os métodos utilizam um subconjunto de janelas, o que pode levar a um resultado subótimo do detector (KING, 2015).

Figura 5 – Exemplo de subjanelas positivas para detecção da face.

Fonte: Foto original por Hannah Nelson obtida de Pexels.

A detecção de objetos por margem máxima (*Max-Margin Object Detection - MMOD*) (KING, 2015) fornece uma otimização para operação em todo o conjunto de subjanelas. O método encontra um subconjunto de janelas que retorna o maior valor ($f\text{-score}$) sem sobreposição simultânea, como exemplificado na figura 6. A minimização dos erros de detecção é realizada selecionando a maior pontuação possível, sendo que a função de perda considera as detecções perdidas e os alarmes falsos.

Figura 6 – Exemplos de subjanelas e seus $f\text{-scores}$. Um método guloso retornaria uma solução subótima com a janela do centro ($f\text{-score}=7$), enquanto a solução ideal para essas três janelas seria as duas do canto ($f\text{-score}=12$).

Fonte: Exemplo inspirado de King (2015) com foto de fundo por Bruce Mars obtida de Pexels.

2.4 Landmarks

Landmarks faciais também conhecidos como *fiducial points* ou *feature points* são pontos localizados principalmente no entorno dos componentes faciais olhos, nariz e boca. (WANG et al., 2017). No contexto do presente trabalho, os *landmarks* são utilizados como delimitadores dos componentes faciais, como exemplificado na figura 7. Existem três abordagens de localização desses pontos referentes aos componentes faciais: manual, semiautomática e automática, como detalhado na seção 3.1. O presente trabalho considera as abordagens automáticas na composição de sua metodologia.

As abordagens de identificação do *landmarks* faciais de forma automática também são denominadas *Facial feature point detection (FFPD)*. Geralmente essas abordagens aprendem, de forma supervisionada, a localização dos *landmarks* faciais com base nos pontos marcados em uma grande quantidade de imagens de forma manual (WANG et al., 2017). Dentre as abordagens de detecção automática analisadas por Wang et al. (2017), as abordagens baseadas em regressão em cascata se destacaram com a implementação das *ensemble regression trees*.

Figura 7 – Exemplo de landmarks faciais para os olhos e a sobrancelha.

Fonte: Rafael Luiz Testa, 2017.

Esta abordagem, *ensemble regression trees*, de identificação dos *landmarks* não é paramétrica, assim ela não se baseia em nenhuma forma de distribuição (WANG et al., 2017). São utilizadas árvores de regressão em cascatas de maneira a aprender os *landmarks* de um determinado componente facial. Na abordagem proposta em Kazemi e Sullivan (2014), os *landmarks* do componente facial são aprendidos com regressão considerando as intensidades dos *pixels*. Para cada nó da árvore, uma decisão é tomada de acordo com um valor de limiar em relação à diferença de intensidades entre dois *pixels* (KAZEMI; SULLIVAN, 2014).

2.5 Função de similaridade

Na seção 5.5 é definida e discutida uma abordagem de seleção das imagens de acordo com a similaridade entre duas imagens. Nela, é preciso verificar, dentro do banco de imagens, quais imagens faciais mais se assemelham à imagem da face na qual se deseja sintetizar a expressão. A busca por essas imagens é realizada considerando as características visuais semelhantes na imagem desejada e na imagem procurada. No caso das imagens faciais, pode-se verificar as faces que possuem os *landmarks* faciais mais semelhantes.

A semelhança entre as imagens pode ser verificada por funções de similaridade. Uma função de similaridade mede a distância entre dois vetores de características retornando um valor não negativo (GONÇALVES; DELAMARO; NUNES, 2017). É comum a utilização da distância Euclidiana como métrica para o algoritmo de função de similaridade (GONÇALVES; DELAMARO; NUNES, 2017).

A distância euclidiana é formulada na equação 1, na qual x é um conjunto de características a ser comparado com o conjunto de características y , e n é a quantidade de características de cada conjunto. Assim, estabelece-se a relação espacial entre os conjuntos de parâmetros.

$$Dist_{Euclid} = \sqrt[2]{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

Outra função de distância que pode ser utilizada como medida de similaridade é a distância de *Canberra* (GONÇALVES; DELAMARO; NUNES, 2017). Ela está formulada na equação 2, na qual x e y são os conjuntos de características e n é a quantidade de características de cada conjunto.

$$Dist_{Canb} = \sum_{i=0}^{n-1} \frac{|x_i| - |y_i|}{|x_i + y_i|} \quad (2)$$

2.6 Triangulação

No processo de sintetização das expressões faciais são necessárias operações nos conjuntos de *landmarks* faciais, de forma a criar triângulos que servirão como base para o alinhamento (Seção 5.4) e a deformação dos componentes faciais (Seção 5.6). A triangulação permite a divisão do componente facial em partes menores a serem processadas.

A triangulação de *Delaunay* é uma maneira de formar triângulos a partir de um conjunto de pontos. Ela otimiza os triângulos gerados maximizando o ângulo mínimo (BERG et al., 2008), ou seja, procura-se obter maiores ângulos internos para os triângulos resultantes. Na figura 8, os pontos das sobrancelhas são triangulados pela técnica de *Delaunay*.

2.7 Transformações afins

No processo de síntese das expressões é necessário o mapeamento das regiões dos triângulos para um novo formato, seja em relação ao alinhamento (Seção 5.4) entre duas imagens ou à deformação dos componentes faciais (Seção 5.6). Para isso, são definidas transformações geométricas que permitem realizar essas operações. No contexto da com-

Figura 8 – Exemplo de triangulação de *Delaunay*.

Fonte: Rafael Luiz Testa, 2017.

putação gráfica, as transformações geométricas permitem o mapeamento dos *pixels* de uma imagem original em outra imagem (modificada), por meio da reorganização espacial ou interpolação das intensidades dos *pixels* (PEDRINI; SCHWARTZ, 2008).

A transformação afim é uma transformação geométrica definida por uma matriz que generaliza as transformações de translação, rotação, escala, espelhamento e cisalhamento (PEDRINI; SCHWARTZ, 2008). A figura 9 mostra um exemplo da transformação da imagem na área interna do triângulo à esquerda para o novo formato de acordo com o triângulo à direita.

Figura 9 – Exemplo de transformação afim.

Fonte: Rafael Luiz Testa, 2017.

2.8 Adequação da iluminação

Um dos passos importantes na geração de cenas realistas é a aplicação dos efeitos relacionados à iluminação nas superfícies visíveis (HEARN; BAKER, 2003). Essa aplicação pode ser realizada por meio de modelos de iluminação. Os modelos de iluminação são usualmente aproximações das leis da física que explicam os efeitos de iluminação nas superfícies (HEARN; BAKER, 2003). Outra maneira de simular os efeitos de iluminação é por meio da obtenção dos efeitos de luz de um ambiente real, que pode ser extraída

de imagens (REINHARD et al., 2005), denominada iluminação baseada em imagem (*Image-Based Illumination*).

A maneira mais comum de realizar essa simulação da iluminação em imagens faciais expressivas é por meio da imagem razão (ver seção 3.2). A iluminação para a nova imagem pode ser produzida pela diferença entre duas outras imagens: uma imagem sobre a iluminação existente e outra sobre a iluminação desejada, sendo ambas da mesma cena (MARSCHNER; GREENBERG, 1997). Ou seja, dadas três imagens, uma com a iluminação desejada (D) na cena A (I_A^D), outra com a mesma cena (A) em outra condição de iluminação (I_A^E) e a terceira com uma cena similar B e uma iluminação existente (E) (I_B^E). Desta maneira, é possível gerar uma nova imagem (I_B^D) com a cena B e a iluminação desejada de I_A^D . Esse procedimento é formulado na equação 3, na qual x e y são as coordenadas nos *pixels* nas imagens.

$$\frac{I_B^D(x, y)}{I_B^E(x, y)} = \frac{I_A^D(x, y)}{I_A^E(x, y)} \quad (3)$$

A equação 3 pode ser reformulada de modo a isolar a nova imagem (I_B^D). Assim, obtém-se a equação 4, na qual a nova imagem é obtida pela multiplicação, *pixel a pixel*, da imagem atual pela razão das imagens da outra cena (MARSCHNER; GREENBERG, 1997). Esta divisão é realizada *pixel a pixel* das imagens em tons de cinza e $I_B^D(x, y)$ é definido como zero quando $I_A^E = 0$.

$$I_B^D(x, y) = I_B^E(x, y) \cdot \frac{I_A^D(x, y)}{I_A^E(x, y)} \quad (4)$$

2.9 Redes neurais convolutivas

As Redes Neurais profundas (*Deep Neural Networks* - *DNN*) são métodos de aprendizado de máquina que descobrem a representação necessária para detecção ou classificação tendo como entrada os dados brutos (LECUN; BENGIO; HINTON, 2015). Nesses métodos as diversas camadas ocultas integram atributos de níveis mais baixos até os níveis mais altos, obtendo novas representações de alto nível (BENGIO, 2009). Desta maneira, essas arquiteturas são comumente usadas para extrair características, como utilizado neste trabalho (ver seção 5.5). As redes neurais convolutivas (*Convolutional Neural Network* - *CNN*) (CUN et al., 1990) são um tipo de rede neural profunda.

Tradicionalmente, as redes neurais são arquiteturas compostas de camadas, em que a camada anterior é conectada a todos os neurônios da camada seguinte (PATTERSON; GIBSON, 2017). A figura 10a mostra um exemplo no qual as camadas são completamente conectadas. Ou seja, todos os elementos da camada de entrada X têm conexão com todos os neurônios da camada oculta H, que por sua vez estão todos conectados à camada de saída Y.

Figura 10 – Entrada para as redes neurais tradicionais e convolutivas.

Fonte: Adaptada de Criscuolo (2017).

Em contrapartida, os neurônios de uma CNN possuem conexões locais com pequenas regiões da entrada (campos receptivos locais). A figura 10b apresenta um exemplo na qual uma região de tamanho quatro por quatro *pixels* (*kernel*) é a entrada para um determinado neurônio. Ou seja, apenas aquela região será responsável pela ativação do neurônio correspondente.

As CNNs são estruturadas como uma combinação de camadas de *pooling*, convoluções e camadas completamente conectadas. A camada convolutiva é composta por um conjunto de filtros que emulam um estímulo visual a um neurônio específico (LECUN; BENGIO; HINTON, 2015) para gerar os mapas de características. A camada de *pooling* combina a saída dos neurônios visando a simplificar e condensar os mapas de características (SCHERER; MÜLLER; BEHNKE, 2010). As camadas completamente conectadas seguem o mesmo princípio de uma rede neural tradicional.

Na primeira parte da arquitetura de uma CNN, a entrada é transformada pelas camadas de convolução e *pooling*. Os mapas de características são ligados à camada anterior

por conjunto de pesos (LECUN; BENGIO; HINTON, 2015). A camada de *pooling* reduz a dimensão dos mapas de características da camada anterior. As camadas finais são redes neurais totalmente conectadas. Esse processo é ilustrado na figura 11.

Figura 11 – Ilustração da arquitetura de uma *CNN*.

2.9.1 Redes Neurais Residuais

Um dos problemas encontrados ao treinar redes neurais profundas é o problema do desaparecimento do gradiente (GLOROT; BENGIO, 2010). Uma solução proposta a este problema foi a Rede Neural Residual (*Residual Neural Network* - ResNet) (HE et al., 2016). Esse problema é resolvido adicionando a identidade de uma camada anterior para uma camada posterior não consecutiva. A ideia é que em vez de aproximar diretamente uma função $H(X)$, aproxima-se uma outra função $F(X)$, tal que $F(X) := H(X) + X$ (ver figura 12) (HE et al., 2016). Ou seja, o objetivo é encontrar a função $F(X)$ que representa a soma ou diferença em relação à identidade.

Figura 12 – Bloco residual em uma ResNet.

Fonte: Adaptada de He et al. (2016).

2.10 Métricas para comparação de imagens

Uma das etapas é a avaliação das imagens sintetizadas. Esta tarefa tem por objetivo verificar se as imagens foram sintetizadas adequadamente. Essa avaliação pode ser realizada por procedimentos subjetivos, com base na análise de observadores humanos, ou objetivos, por meio da verificação de similaridade ou diferenças entre duas imagens (PEDRINI; SCHWARTZ, 2008). Uma análise mais completa dessas abordagens é apresentada na seção 3.5.

As métricas objetivas podem ser divididas em três tipos: referência completa (*Full-Reference*), referência reduzida (*Reduced-Reference*) e sem referência (*No-Reference*) (WANG; BOVIK, 2006). Para este trabalho considerou-se apenas as métricas de referência completa porque existe uma imagem facial expressiva disponível para comparação.

Assim, com o uso de métricas de referência completa, a avaliação das imagens sintetizadas pode ser feita por meio da comparação entre a imagem processada e uma imagem referência. Foi observado na seção 3.5 que métricas baseadas no erro quadrático médio (*Mean Squared Error - MSE*) foram as mais utilizadas no contexto da avaliação da síntese de expressões faciais.

O *MSE* e o *Average Peak Signal to Noise Ratio* (PSNR) são as métricas de diferença mais populares na avaliação do processamento de imagens e vídeos (WINKLER, 2005). O *MSE* é um estimador de erro baseado na diferença quadrática entre os valores dos *pixels* em duas imagens, conforme apresentado na equação 5, na qual I é a imagem avaliada, R é a imagem referência, m é altura em *pixels* e n é largura em *pixels*, assim, x e y são as coordenadas dos *pixels*.

$$MSE = \frac{1}{mn} \sum_{x=1}^m \sum_{y=1}^n [I(x, y) - R(x, y)]^2 \quad (5)$$

No contexto de processamento de imagens, o PSNR é definido pela razão entre o maior valor possível para um *pixel* e o ruído de distorção que afeta sua qualidade. Sua formulação é apresentada na equação 6, na qual max é o máximo valor possível para o *pixel* e *MSE* está definido na equação 5. Enquanto o *MSE* mede a imagem diferença, o PSNR mede a fidelidade da imagem em relação à imagem referência (WINKLER, 2005).

$$PSNR = 10 \log_{10} \frac{max^2}{MSE} \quad (6)$$

2.11 Considerações finais

Os conceitos descritos neste capítulo têm por objetivo auxiliar na compreensão dos demais capítulos, sendo principalmente usados: *CNN* para detecção facial (Seção 5.1), os *landmarks* para correspondência dos elementos faciais (Seção 5.2), medidas de similaridade e *ResNet* para seleção das imagens fonte (Seção 5.5), transformações geométricas para alinhamento das faces e deformação (Seções 5.4 e 5.6) e adequação da iluminação por meio do mapeamento de iluminação (Seção 5.7).

Também existem conceitos relacionados a outros aspectos do trabalho além do processo de síntese de expressões faciais, sendo eles: as métricas de avaliação comparativa que são usados na avaliação dos resultados (Seção 4.2.2) e as origens das expressões faciais que ajudam a limitar o escopo do trabalho. Essa limitação das expressões faciais está relacionada tanto às emoções não sentidas (comunicação não verbal), quanto às emoções sentidas (estados mentais), apresentados na figura 1.

3 Levantamento bibliográfico

Este capítulo apresenta os principais pontos de uma Revisão Sistemática (RS) conduzida com objetivo de identificar os métodos e técnicas existentes na literatura para sintetizar e avaliar expressões faciais em fotografias. A lista completa dos artigos incluídos, assim como discussões mais detalhadas são apresentadas em Testa et al. (2018a).

A presente RS foi executada seguindo as três fases clássicas: planejamento, seleção de trabalhos; extração e interpretação dos dados (PETTICREW; ROBERTS, 2008).

Buscou-se a *string* (“*facial expression*” AND (*synthesis OR map*)) AND NOT (*3D OR robot*)) em título, resumo e palavras-chaves dos artigos presentes nas bases acadêmicas *ScienceDirect*¹, *IEEE Xplore*² e *ACM Digital Library*³, tendo sido retornado 437 artigos. Nesses artigos foram avaliados os critérios de inclusão e exclusão apresentados na tabela 1 com base na leitura de título, palavras-chave e resumo. Ao final, 53 artigos foram incluídos para análise completa.

Tabela 1 – Critérios de inclusão e exclusão

Critério	Descrição
I1	Serão incluídos trabalhos que abordem a síntese completa ou parcial de expressões faciais em imagens.
I2	Serão incluídos trabalhos que descrevem métodos de avaliação em relação a síntese de expressões faciais.
E1	Serão desconsiderados trabalhos em que a síntese ocorre em imagens não bidimensionais.
E2	Serão desconsiderados trabalhos em que a síntese não ocorre em fotos.
E3	Serão desconsiderados trabalhos em que a síntese ocorre em vídeos ou utilize vídeos em sua metodologia.
E4	Serão desconsiderados trabalhos que abordam apenas a síntese de fala.
E5	Serão desconsiderados trabalhos que utilizam sensores de movimento no processo síntese das imagens.
E6	Serão desconsiderados trabalhos duplicados.

A qualidade dos trabalhos incluídos foi medida a partir da avaliação de critérios de inclusão relacionados: características das imagens fonte, abordagem de identificação e especificação dos *landmarks* faciais, aspectos das imagens sintetizadas e descrição das

¹ <http://www.sciencedirect.com/>

² <http://ieeexplore.ieee.org/>

³ <http://dl.acm.org/>

Tabela 2 – Critérios de qualidade separados por fator e porcentagem de atendimento aos mesmos.

Fator	Critério	Descrição	Atendimento ao critério
Imagens fonte	Q1	O estudo especifica a base de imagens utilizada ou as características das imagens adquiridas.	72% de 53
Landmarks faciais	Q2	O estudo especifica a forma de identificação dos <i>landmarks</i> faciais.	74% de 47
	Q3	O estudo especifica quais <i>landmarks</i> faciais são utilizados.	74% de 47
Imagens sintetizadas	Q4	As imagens sintetizadas pelo estudo apresentam a região interna da boca (dentes, lábios e língua).	78% de 51
	Q5	As imagens sintetizadas pelo estudo apresentam enrugamentos.	57% de 53
	Q6	O estudo mostra pelo menos um exemplo das imagens sintetizadas.	96% de 53
Avaliação Objetiva dos resultados	Q7	O estudo apresenta uma avaliação objetiva da síntese de expressões faciais.	26% de 53
	Q8	A avaliação objetiva dos resultados compara as imagens sintetizadas com as imagens referência (<i>ground truth</i>).	25% de 16
	Q9	A avaliação objetiva dos resultados compara as imagens sintetizadas com as imagens sintetizadas por outros trabalhos.	44% de 16
Avaliação subjetiva dos resultados	Q10	O estudo apresenta uma avaliação subjetiva da síntese de expressões faciais.	19% de 53
	Q11	A avaliação subjetiva dos resultados apresenta as características do usuário.	31% de 13
	Q12	A avaliação subjetiva dos resultados apresenta algum mecanismo de coerência.	31% de 13

avaliações objetiva e subjetiva, conforme apresentado na tabela 2. A qualidade média desses aspectos é ilustrada na figura 13. Enquanto os critérios referentes aos aspectos das imagens foram encontrados em uma maior quantidade de trabalhos, aspectos relacionados à avaliação dos resultados estiveram menos presentes.

Foram encontradas abordagens baseadas em Deformação e Mapeamento (Seção 3.2), Aprendizado de Máquina (Seção 3.3), e outras (Seção 3.4). Algumas dessas abordagens

Figura 13 – Gráfico da porcentagem média do atendimento aos critérios de qualidade.

Fonte: Adaptado de Testa et al. (2018a).

misturam técnicas de duas abordagens diferentes e foram classificadas como híbridas. A distribuição dos trabalhos nessas abordagens é apresentada na figura 14. A figura 15 ilustra as técnicas utilizadas em cada etapa pelos artigos para compor uma abordagem (pré-processamento, síntese e pós-processamento), e são detalhadas nas seções 3.2 a 3.4.

Figura 14 – Diagrama de *Venn* da distribuição das abordagens.

Fonte: Adaptado de Testa et al. (2018a).

3.1 Pré-processamento

Vários trabalhos executam etapas de pré-processamento, as quais independem da abordagem de sintetização adotada *a posteriori*. Essas etapas podem ser divididas em

Figura 15 – Estágios da síntese de expressões faciais por abordagem. Linhas tracejadas indicam um estágio opcional. Retângulos com linhas tracejadas indicam estágios opcionais e setas com linhas tracejadas indicam que apenas um dos caminhos deve ser seguido.

Fonte: Adaptado de Testa et al. (2018a).

identificação dos *landmarks* (correspondência de características) faciais, alinhamento das faces e seleção das imagens fonte, conforme detalhado a seguir.

Landmarks faciais: pontos em posições específicas da face permitem realizar a correspondência entre as características de duas faces diferentes. Os *landmarks* faciais foram marcados manualmente em 38% (20) dos estudos, semiautomáticas em 8% (4), automáticas em 17% (9) e retirados do banco de dados de expressões faciais em 4% (2) dos estudos. Dos demais estudos, 19% (10) não especificaram e 15% (8) não extraíram pontos de referência. A maior parte dos trabalhos utilizou entre 60 e 80 pontos, considerando olho, boca, nariz e contorno do rosto.

Alinhamento das faces: as faces podem apresentar diferentes proporções e posicionamento devido às diferenças faciais e à aquisição das imagens. O alinhamento é

o pré-processamento realizado para ajustar essas diferenças e garantir melhores resultados. Comumente consiste em duas etapas: a triangulação e a transformação. A triangulação mais aplicada foi a de *Delaunay*. Os métodos para transformação das imagens foram *Affine Transformation* (15% - 8), *Procrustes Transformation* (4% - 2), uma combinação de *Procrustes Analysis* com *Affine Transformation* (6% - 3), *pixel-based correspondence* (2% - 1), *Thin Plate spline (TPS)* (2% - 1), e normalização com FAPU (facial animation parameters unit) (2% - 1).

Seleção das imagens fonte: nesta etapa os trabalhos selecionam quais imagens fonte da base serão usadas para a síntese da expressão facial desejada. Sessenta e seis por cento (35) dos trabalhos analisados usam todas as imagens da base para composição do modelo, 28% (15) dos trabalhos não especificaram como é realizada a escolha das imagens fonte e, por fim, os demais estudos apresentaram medidas de similaridade para seleção das imagens.

As medidas de similaridade foram divididas em dois grupos: similaridade baseada na expressão facial e similaridade baseada na estrutura. A primeira busca a expressão facial desejada em imagens da face destino, enquanto a segunda visa encontrar faces com características parecidas. Em Xiong et al. (2007) e Xiong et al. (2010) encontra-se o olho mais próximo utilizando a distância euclidiana como medida de similaridade baseada em estrutura, porém este trabalho se restringe à síntese da expressão facial para os olhos. Mohammadian, Aghaeinia e Towhidkhah (2016) busca as cinco faces mais similares (*KNN*) utilizando a estrutura de toda a face (*landmarks*), então a média desses cinco vizinhos próximos compõem o modelo.

3.2 Síntese baseada em Deformação e Mapeamento

“Deformação e Mapeamento” é a primeira abordagem a ser apresentada para síntese de imagens faciais expressivas. Para a síntese da nova expressão facial usualmente é necessária uma imagem para F_{fn} , uma para F_{fe} e uma para F_{dn} . A seguir são descritas as principais abordagens para a etapa de deformação e mapeamento:

Deformação: nesta etapa os *landmarks* da F_{dn} são alterados de acordo com as faces fontes (F_{fn} e F_{fe}) para corresponder à face destino expressiva. A face é, então, dividida em partes menores, usualmente triângulos. Por fim, os triângulos da face

destino neutra são deformados por meio de transformações geométricas para a face destino expressiva. Para isso, os trabalhos analisados utilizaram abordagens como: triangulação e *warping* (21% - 11), *Spline warping* (11% - 6), tabelas de deformação facial com deformações quadráticas (2% - 1), deformação baseada na preservação das bordas (2% - 1) e os demais (8% - 4) não especificaram a abordagem de deformação.

Mapeamento: o mapeamento das informações de iluminação permite sintetizar aspectos mais finos como os enrugamentos. A ideia é transferir as mudanças de iluminação que ocorrem entre as imagens fonte expressiva e neutra. Para isso, em 21% (11) dos trabalhos *Expression Ratio Image (ERI)*, 6% (3) usou *Lighting Differences*, 1% (2) usou *Blend ratio*, 2% (1) usou *Helmholtz-Hodge decomposition (HHD)* e outros 2% (1) análise de frequência. Oito por cento (4) não mapearam as características de iluminação.

3.3 Síntese baseada em Aprendizado de máquina

Na abordagem baseada em aprendizado de máquina a expressão facial desejada é aprendida com base em múltiplas instâncias dessa expressão. Para isso, um modelo é construído para síntese das expressões faciais a partir das características relevantes de cada imagem do conjunto das faces fonte e combinação dessas características, como descrito a seguir.

Características faciais extraídas: as características extraídas foram categorizadas em três tipos: forma, textura e iluminação. Para a primeira abordagem são consideradas as alterações nas estruturas faciais: *landmarks* e *patches* das faces. Para a segunda abordagem os trabalhos consideram os posicionamentos e relações entre os próprios *pixels* da imagem como características a serem aprendidas. Por fim, a terceira trata das informações de cor e brilho da imagem para construção do modelo. Há também abordagens que utilizam uma combinação dessas características extraídas.

Redução de dimensionalidade: os trabalhos realizaram a redução de dimensionalidade com métodos de fusão. Foram utilizados *Principal Component Analysis (PCA)*, *Statistical Shape and Appearance Models*, *Singular Value Decomposition (SVD)*,

Supervised Locality Preserving Projections (SLPP) e *encoders* de faces com redes neurais.

A maior parte dos trabalhos baseados em aprendizado de máquina utilizaram o aprendizado supervisionado (64% - 34) e os demais (6% - 3) o aprendizado não supervisionado.

Aprendizado supervisionado: os métodos supervisionados utilizados para o treinamento do modelo foram: Regressão, Máquinas de Vetores-Suporte, Redes Neurais e Redes Neurais Profundas.

Aprendizado não-supervisionado: agrupamento com correlação e *Generative Adversarial Networks* foram as duas abordagens encontradas para construção de modelos de aprendizado não supervisionado.

3.4 Sintetização com outras abordagens

Foram enquadradas em outras abordagens os artigos encontrados em menores quantidades que apresentaram uma solução diferente para o problema. Foram encontradas as abordagens de *Blend Shape*, simulação da musculatura, *cloud model*, e baseada em expressão facial similar.

3.5 Avaliação dos resultados

As métricas de avaliação encontradas nos estudos incluídos foram divididas em objetivas e subjetivas. As métricas objetivas são aquelas que avaliam quantitativamente aspectos das imagens sintéticas. Já as métricas subjetivas são aquelas em que as pessoas avaliam o desempenho do sistema.

Métricas objetivas as imagens sintetizadas foram comparadas com outras imagens, tais como imagens *ground truth* (comparar com o esperado), imagens sintetizadas por outros trabalhos (comparar diferentes abordagens) e imagens fonte expressivas (verificar identidade). Para essa comparação, os trabalhos incluídos utilizaram: *Mean Square Error (MSE)*, correlação, robustez, similaridade estrutural, consistência dos *landmarks*, e reconhecimento da expressão facial. Também foi encontrada uma

abordagem que não utiliza imagens referência (comparação de duas imagens), o *Q-value*. Essa abordagem avalia aspectos de bloqueio (*blocking*) e desfoque (*blurring*) na imagem sintética.

Métricas subjetivas nessa abordagem as imagens são avaliadas pela percepção das pessoas. Eles avaliaram aspectos como: o realismo das imagens, a expressão facial, e a preservação da identidade. Além das imagens sintetizadas também foram utilizadas imagens de outros trabalhos, *ground truth*, e diferentes imagens de um indivíduo.

3.5.1 Avaliação dos resultados pelos autores

A disponibilização de imagens em diferentes condições, comparação com imagens sintetizadas por outras abordagens, comparação com imagem *ground truth* e enrugamentos foram os aspectos comumente discutidos nos trabalhos incluídos. Os estudos se preocuparam em gerar diferentes expressões faciais, diferentes faces/pessoas, variações nas intensidades das expressões faciais, enrugamentos, expressões faciais mistas, transição entre duas expressões faciais e a preservação de identidade. Também foram relatados testes considerando diferentes condições de iluminação, melhoria na precisão da detecção *landmarks* faciais, desfoque (*blurriness*), configurações inválidas para uma expressão facial, oclusão, variação de pose, síntese em faces não humanas (como desenhos e pinturas), e tempo de execução.

3.6 Bancos de imagens de expressões faciais

Os bancos de imagens de expressões faciais mencionados nos estudos incluídos nesta RS contêm as imagens utilizadas como base para sintetizar expressões faciais (imagens fontes) e servem para avaliar as abordagens propostas pelos autores. Dos 12 bancos de expressões faciais analisados, destacam-se *CK/CK+* (KANADE; COHN; TIAN, 2000; LUCEY et al., 2010) e *JAFFE* (LYONS et al., 1998) por serem mais utilizados. Em geral, os bancos de imagens contêm aspectos importantes para avaliação das imagens geradas de acordo com os aspectos analisados pelos autores (ver seção 3.5.1), sendo eles: variadas expressões faciais, mesma face neutra e expressiva, variados tipos de enrugamentos, exemplos de diferentes pessoas, variadas intensidades para uma mesma expressão facial, e diferentes

condições de iluminação. Apesar disso, alguns aspectos importantes ficaram a desejar em alguns bancos de expressões faciais, por exemplo, alguns bancos não apresentaram as seis expressões faciais consideradas universais (ver capítulo 1), outros contêm um número muito pequeno de indivíduos (10 a 18), e alguns não contêm variação de etnia e gênero (JAFPE).

3.7 Discussões

As seções a seguir apresentam as principais reflexões tecidas a partir da análise do conjunto de artigos incluídos nesta RS. Além da análise das tendências dos artigos incluídos na RS, são vislumbrados desafios e algumas formas de contorná-los, as quais podem ser indicativos de oportunidades de pesquisa.

3.7.1 Métodos de pré-processamento

Seleção das imagens fonte: muitas vezes não está disponível imagens de faces expressivas da pessoa destino nos bancos. Por isso, a abordagem de seleção das imagens fonte baseada na expressão facial similar só é possível para algumas aplicações. Já no caso da abordagem de similaridade baseada em estrutura, é necessário um banco de imagens que possua uma variedade de etnias, faixas etárias e de ambos os gêneros para aumentar as chances de encontrar uma face realmente similar, o que não pode ser encontrado em todos os bancos de imagens descritos.

Identificação dos landmarks faciais: para evitar a preocupação com sistemas de detecção de *landmarks* faciais com precisão, a maior parte dos trabalhos fez a identificação desses pontos de forma manual. Entretanto, a abordagem manual é trabalhosa em um contexto em que há um grande número de imagens a serem processadas. Outro aspecto importante da identificação dos *landmarks* faciais que influencia nas imagens sintetizadas são as regiões consideradas. Embora a maior parte dos trabalhos tenha identificado *landmarks* nas principais regiões da face (olhos, boca, sobrancelhas, nariz, contorno do rosto), apenas alguns consideram componentes faciais como bochecha e enrugamentos, o que poderia melhorar o realismo das imagens sintetizadas.

Alinhamento das faces: o alinhamento das faces é importante para que as imagens possuam uma melhor correspondência dos componentes faciais. Uma correspondência mais apurada consegue mapear melhor as características das faces fonte para a face destino, tanto para deformação e mapeamento, quanto para construção de um modelo mais robusto. Uma alternativa seria o alinhamento por correspondência densa, por exemplo, utilizando *SIFT Flow* (LIU; YUEN; TORRALBA, 2011).

3.7.2 Métodos de síntese

As abordagens de “Aprendizado de Máquina” têm a vantagem de definir os aspectos que caracterizam a expressão relativa a uma emoção, porque elas sintetizam as características expressivas com base em mais de uma instância da mesma expressão facial. Porém, ela exige uma alta carga de processamento para construção do modelo, e também um número grande de instâncias para resultados mais realistas, especialmente nas abordagens de aprendizado profunda (ZHOU; SHI, 2017). Isso pode ser problemático, principalmente no caso de expressões faciais especiais em que há poucas ou apenas um exemplo da expressão desejada (XIONG et al., 2010). É conhecido na literatura (EKMAN; ROSENBERG, 1997) que algumas emoções são difíceis de serem reproduzidas de forma espontânea e, por isso, são mais difíceis de serem representadas em bases de imagens. Muitas vezes, quando representadas, é pedido para o sujeito simular a expressão facial (KANADE; COHN; TIAN, 2000), o que pode gerar uma expressão artificial e, consequentemente, influenciar no realismo desejado para a imagem destino.

Por outro lado, as abordagens de “Deformação e Mapeamento” (Seção 3.2) conseguem reproduzir qualquer expressão facial mesmo com apenas um exemplo. Porém, as diferenças entre as faces fonte e destino podem tornar a transferências das características expressivas inviáveis porque não há casamento aceitável entre os componentes das duas imagens. Em consequência, o resultado pode ser extremamente superficial. Alguns estudos juntam as técnicas de deformação, mapeamento e aprendizado para compor sua abordagem, tentando fazer o melhor de cada abordagem para sintetizar a nova imagem.

A partir do cenário analisado, verifica-se que ainda é um desafio o desenvolvimento de abordagens híbridas, que sejam capazes de agregar pontos fortes das abordagens clássicas. Seria interessante o desenvolvimento de abordagens que possibilitassem a síntese de

expressões considerando uma maneira adaptativa de utilizar as imagens fonte, considerando a quantidade e os tipos de imagens disponíveis para a reconstituição (síntese) da expressão facial.

Outro problema em aberto é a síntese das expressões faciais em ambientes não controlados, por exemplo, com diferentes condições de iluminação, e diferentes poses. No caso da iluminação alguns trabalhos tentam criar modelos para reduzir a influência das variações de iluminação (XIONG et al., 2010; ZHANG et al., 2014; XIE; SHEN; JIANG, 2017). Tais modelos são desenvolvidos principalmente porque a abordagem de mapeamento mais utilizada, a *ERI*, é sensível às variações de iluminação.

3.7.3 Métricas de avaliação dos resultados

Dos dois tipos de métricas analisadas, as avaliações objetivas foram as mais utilizadas. A avaliação objetiva é realizada de forma sistemática e quantitativa, usando métricas que não dependem da percepção de uma pessoa para ser executada. Desta maneira, consegue-se tornar a avaliação replicável e de fácil comparação com outros trabalhos. Portanto, essas métricas, além de avaliar os resultados, podem servir como base para a comparação entre estudos. A avaliação subjetiva avalia os resultados por meio do questionamento da percepção do usuário sobre as imagens sintetizadas. Essa percepção é importante para garantir que as imagens sintetizadas são convincentes aos olhos de quem as utiliza. Apesar de relativamente simples, nesta abordagem existe o desafio de alcançar uma quantidade de usuários significativa para avaliar os resultados, além da necessidade de construir experimentos adequados, que atinjam o objetivo sem causar exaustão ao usuário. Não foram percebidos nos artigos analisados a descrição sistemática de protocolos de experimento. Assim, o estudo de protocolos para condução de experimentos poderia contribuir para um avanço no estado da arte desta área de pesquisa.

A criação de métricas e protocolos para avaliação dos resultados exige a padronização e disponibilização de alguns aspectos. É preciso definir o conjunto de métricas de avaliação importantes para avaliar cada aspecto. Também seria interessante a disponibilização das imagens sintetizadas para fins comparativos. A exemplo de outras áreas, como Engenharia de Software (SJOEBERG et al., 2005; ANDREWS; BRIAND; LABICHE, 2005; THELIN; RUNESON; WOHLIN, 2003), a criação de repositórios para disponibilizar programas e

benchmarks para avaliação de novos métodos é um desafio que pode se transformar em uma importante contribuição para a pesquisa de síntese de imagens faciais. Para uma avaliação mais completa, as métricas objetivas e subjetivas poderiam ser empregadas de forma complementar, criando-se uma abordagem que considere tanto a percepção do usuário quanto a quantificação dos resultados.

3.7.4 Bancos de imagens de expressões faciais

Apesar dos bancos de imagens de expressões faciais possuírem muitos dos aspectos necessários para a síntese das expressões faciais, ainda restam alguns desafios que permitiriam alcançar melhores resultados. Por exemplo, de todos os bancos de expressões faciais analisados apenas um continha expressões faciais espontâneas. Isso porque essas imagens são difíceis de coletar e atribuir emoção. Outro aspecto importante para os bancos de imagens seria a aquisição de um número maior de imagens, principalmente com o aumento no número de trabalhos que utilizam o aprendizado profundo.

3.8 Conclusões

Neste estudo foi conduzida uma pesquisa bibliográfica sistemática com o intuito de analisar as abordagens existentes para síntese de expressões faciais em fotografias, assim como elementos adicionais que influenciam nos resultados dessas abordagens: as técnicas de pré-processamento, os bancos de expressões faciais e as diferentes formas e métricas de avaliação dos resultados. A síntese de expressões faciais é um tópico desafiador. Uma abordagem precisa deve possibilitar a geração da nova expressão facial de forma consistente, realista e que preserve a identidade da face. Apesar de ser uma área relativamente bem explorada na literatura, alguns desafios ainda permanecem, tais como: tratamento de oclusões, variações de pose, variações das condições de iluminação, inexistência de padrões e protocolos de avaliação dos resultados. Também é importante a criação de protocolos para avaliação dos resultados encontrados com métricas objetivas e subjetivas.

4 Materiais e métodos

Este capítulo apresenta os materiais, procedimentos e fases inerentes ao desenvolvimento do método proposto (Capítulo 5) e a sua validação (Seção 4.2). As fases abrangidas por este estudo são descritas nas seguintes seções: a seção 4.1 indica as técnicas que compõem o método proposto, a seção 4.2.1 explicita os bancos de expressões faciais utilizados e as suas características, a seção 4.2.2 descreve as métricas de avaliação objetiva e a seção 4.2.3 expõe a métrica de avaliação subjetiva. Além disso, o capítulo 3 ressalta aspectos importantes da literatura para definição do método proposto e definição do banco de expressões. As relações entre essas diversas fases do projeto são ilustradas na figura 16.

Figura 16 – Relação entre as diferentes fases do projeto.

Fonte: Rafael Luiz Testa, 2017

4.1 Desenvolvimento do método proposto

Nesta seção são apresentados os motivos para escolha da abordagem e das técnicas adotadas para o desenvolvimento do método proposto. A descrição do funcionamento do método proposto é apresentada no capítulo 5.

No capítulo 3 as abordagens foram divididas em quatro categorias de acordo com o método utilizado. Dentre as abordagens apresentadas, a abordagem de “Deformação e mapeamento” (Seção 3.2) é a utilizada neste trabalho. Esta abordagem foi escolhida por viabilizar a síntese de uma maior variedade de expressões faciais, mesmo as menos comuns, porque ela exige apenas um exemplo da expressão facial desejada para sua

reconstituição. Apesar de o presente trabalho considerar as seis emoções comumente consideradas universais, o método proposto pode ser empregado para qualquer expressão facial.

De forma geral as abordagens baseadas em “Deformação e Mapeamento” consistem na seleção da imagem fonte, correspondência das características faciais (identificação dos *landmarks* faciais), alinhamento das faces, deformação e mapeamento; como descrito no capítulo 3 e esquematizado na figura 15.

4.1.1 Seleção das imagens fonte

A primeira etapa de seleção das imagens fonte é realizada buscando considerar as particularidades na face de cada indivíduo, porque expressões faciais tendem a ser diferentes entre dois indivíduos. Na maior parte dos trabalhos de “Deformação e Mapeamento” a seleção das imagens fonte é realizada de forma manual. Essa forma de seleção assegura uma imagem fonte adequada, entretanto, ela torna a abordagem dependente de uma etapa manual, o que pode ser inviável para um número grande de imagens.

Alguns trabalhos propõem uma seleção automatizada das imagens fonte, sendo elas: seleção baseada na expressão facial e seleção baseada na estrutura (ver Seção 3.1). Enquanto a primeira exige imagens expressivas da pessoa destino, a segunda busca nas imagens fonte disponíveis a que mais se assemelha à face destino. A abordagem proposta neste trabalho para seleção das imagens fonte se assemelha mais à seleção baseada na estrutura.

Embora a abordagem proposta se assemelhe à seleção baseada na estrutura, ela inclui alguns aspectos que a tornam mais completa e robusta. É mais completa porque considera a face inteira e não apenas a região do olho como em Xiong et al. (2007), Xiong et al. (2010). Além disso, ela busca resultados melhores ao extrair características relativas à identidade da face (ver seção 5.5), enquanto os trabalhos encontrados utilizam diretamente os *landmarks* faciais (XIONG et al., 2007; XIONG et al., 2010) ou *pixels* da imagem (MOHAMMADIAN; AGHAEINIA; TOWHIDKHAH, 2016).

4.1.2 Detecção facial

A detecção facial é uma etapa necessária para identificar e delimitar a região de uma face na imagem, sendo essencial porque determina a região da imagem que será utilizada nas demais etapas.

Para esta etapa foi escolhida uma abordagem por meio de modelos *CNN* porque estas abordagens possuem um bom desempenho para detecção de faces mesmo em condições variadas de pose e iluminação (ZAFEIRIOU; ZHANG; ZHANG, 2015).

4.1.3 Correspondência entre características faciais

A forma mais comum de encontrar a correspondência entre as características faciais em imagens diferentes é por meio da detecção dos *landmarks* faciais. Ela pode ser feita de forma manual ou automática. A abordagem automática foi escolhida por facilitar a identificação para um número grande de imagens, além de automatizar o processo.

Dentre as várias abordagens de detecção automática dos *landmarks* faciais a escolhida foi a *ensemble regression trees* (Seção 2.4), devido ao desempenho superior apresentado em relação a outras abordagens do estado da arte (WANG et al., 2017).

4.1.4 Alinhamento das imagens faciais

O alinhamento das imagens faciais é composto por duas etapas: padronização das posições dos *landmarks* e transformação da imagem.

Nenhum dos trabalhos encontrados na literatura descreve a técnica utilizada na primeira etapa. No presente trabalho, a primeira etapa é realizada de acordo com as posições da pupila (Seção 5.4). Para a segunda etapa a imagem é transformada por transformações afins.

4.1.5 Deformação e Mapeamento

Para as etapas de deformação e mapeamento foram utilizadas as abordagens mais citadas na literatura: transformações afins para a deformação da face e a *ERI* para o mapeamento da iluminação.

Adicionalmente foram realizadas melhorias em alguns aspectos do processo básico utilizando estas técnicas. Essas melhorias e a descrição completa são especificadas no capítulo 5.

4.2 Avaliação dos resultados

A avaliação dos resultados do método proposto é composta de métricas objetivas (Seção 4.2.2) e subjetivas (Seção 4.2.3). Essas avaliações utilizam imagens de bancos de expressões faciais públicos (Seção 4.2.1).

Os resultados gerados a partir das diferentes métricas de avaliações aplicadas aos bancos de expressões faciais são apresentados no capítulo 6.

4.2.1 Bancos de expressões faciais

A primeira etapa para avaliação dos resultados é definir os bancos de imagens de expressões faciais. Alguns aspectos foram avaliados nos bancos de expressões faciais encontrados para seleção dos que mais se adequassem para a avaliação dos resultados

imagens: esses bancos deveriam possuir imagens frontais da face de pessoas encenando ou expressando as seis emoções comumente consideradas universais. Neste aspecto também é importante que os bancos forneçam informações sobre a emoção relacionada à imagem.

disponibilidade: foram utilizados bancos que estivessem publicamente disponíveis *online* ou por requisição, como formulário ou *e-mail*. Foram considerados apenas bancos gratuitos.

diversidade nas características faciais: uma das etapas do método é a busca por imagens faciais similares (Seção 4.1.1). Para garantir um teste mais completo da

busca, as imagens fonte devem possuir características faciais diversas, como faces de várias etnias, faixas etárias e de ambos os gêneros.

Os experimentos foram conduzidos em três bancos de expressões faciais: IMPA FACES (MENA-CHALCO; JUNIOR; VELHO, 2008), *Multimedia Understanding Group* (MUG) (AIFANTI; PAPACHRISTOU; DELOPOULOS, 2010) e *LFW* (*Labeled Faces in the Wild Home*). As principais características desses bancos de imagens são apresentadas na tabela 3.

Tabela 3 – Características dos bancos de imagens utilizados nos experimentos.

Banco de expressões faciais	Expressões faciais	Número de imagens	Número de sujeitos	Resolução
IMPA Faces	Neutra, tristeza, surpresa, satisfação, medo, raiva, aversão, aberto, fechado, beijo, lado esquerdo, lado direito, sagital esquerdo, sagital direito, nuca e testa.	534 imagens	86	640x480
MUG Faces	Neutra, tristeza, surpresa, satisfação, medo, raiva e aversão.	1462 sequências	200	896x896
LFW	Não se aplica.	13233 sequências	5749	150x150

Os bancos de imagens IMPA FACES e MUG são bancos de expressões faciais em que as faces fornecidas expressam as seis emoções faciais consideradas universais, e tais expressões são utilizadas tanto para a síntese da nova imagem quanto como imagem referência (*ground truth*). Cada banco é testado separadamente dos demais. O banco *LFW*, apesar de não possuir expressões faciais rotuladas, contém um grande número de sujeitos o que o torna adequado para o experimento das faces similares.

As imagens em comum desse banco (*LFW*) em relação aos bancos usados para de treinamento, isto é, *VGG face dataset* (PARKHI; VEDALDI; ZISSERMAN, 2015) e *FaceScrub* (NG; WINKLER, 2014) (ver seção 5.5), foram excluídas dos testes. Além disso, foram considerados apenas os indivíduos que possuíam pelo menos duas imagens (1680 sujeitos).

No caso do MUG as imagens são disponibilizadas como *frames* de um vídeo. Assim, para este banco de imagens, foi necessário selecionar manualmente um *frame* com a face neutra e um *frame* com uma face expressiva para cada uma das seis emoções.

4.2.2 Métricas objetivas

Três métricas objetivas foram aplicadas para avaliar diferentes aspectos das imagens sintetizadas pelo método proposto, sendo elas: avaliação da busca por face similar (Seção 6.3), síntese com face similar (Seção 6.4) e preservação da identidade (Seção 6.5).

O principal objetivo da avaliação é verificar se a seleção de uma face fonte mais similar à face destino resulta em um maior realismo das expressões faciais sintetizadas, tendo em vista que esta etapa configura o maior diferencial do trabalho. Sendo assim, para cada uma das métricas objetivas as imagens sintéticas são avaliadas em três cenários diferentes: aleatório, similaridade baseada em cor e similaridade baseada em características faciais.

Experimento para a avaliação da busca por face similar

A avaliação da busca por face similar tem como objetivo verificar se o método é capaz de encontrar a face mais similar à face destino. Cada um dos três diferentes métodos (aleatório, similaridade baseada em cor e similaridade baseada em características faciais) retorna uma imagem que corresponde à imagem mais similar. O método baseado em características faciais está descrito na seção 5.5, enquanto, as abordagens comparadas para efeito de avaliação são descritas a seguir.

Similaridade baseada em cor: consiste em avaliar a similaridade de duas faces por meio da comparação entre os histogramas de cores da região interna da face entre às duas imagens. A região interna da face é cortada de acordo com os *landmarks* do contorno da face e da sobrancelha.

Aleatório: uma imagem fonte neutra é escolhida aleatoriamente entre as imagens do banco de expressões faciais. Desta maneira, a face sintética varia para cada execução.

É verificado se a imagem mais similar pertence à uma imagem do mesmo indivíduo. Para cada teste é pontuado um caso seja a mesma pessoa e zero caso seja outra pessoa. Esta operação é repetida até que todas as imagens do mesmo indivíduo sejam recuperadas. Assim, a equação 7 mostra o cálculo da precisão, sendo VP (verdadeiro positivo) o número de imagens recuperadas até o momento que pertencem à mesma pessoa e $IRec$ o número de imagens recuperadas até o momento. A equação 8 formula o cálculo da revocação, sendo que $IRel$ é o número total de imagens da mesma pessoa.

$$p = \frac{VP}{IRec} \quad (7)$$

$$r = \frac{VP}{IRel} \quad (8)$$

A primeira face retornada deve ser sempre a mesma imagem referência, exceto para a busca aleatória. O experimento foi avaliado no banco de imagens *LFW*. Este banco foi escolhido por conter um grande número de sujeitos com faces em diversas condições (*in the wild*). Os resultados do experimento descritos nesta seção são apresentados na seção 6.3.

Experimento para avaliação da síntese com face similar

O principal objetivo deste experimento é avaliar se a face sintética se aproxima da face referência (*Ground truth*).

Primeiramente, seleciona-se a face selecionada neutra (FSeN) e a face selecionada expressiva (FSeE) desempenhada pelo mesmo indivíduo. Em seguida, sintetiza-se à expressão desejada na face neutra (FSeN), utilizando-se a abordagem proposta pelo presente projeto, gerando-se a face sintetizada expressiva (FSiE). Então, compara-se a similaridade entre a FSeE e a FSiE. Nesta comparação avalia-se a quão próximas as imagens sintetizadas estão em relação às imagens *ground truth* (FSeE). Para este experimento foram escolhidos os bancos de expressões faciais IMPA e MUG .

Os resultados referentes a este experimento são apresentados na seção 6.4.

Experimento para avaliação da preservação da identidade

O principal objetivo desta métrica é avaliar se características faciais da face foram preservadas após a alteração da expressão facial. Para isso, um processo similar ao explicado na avaliação da síntese com face similar (Seção 4.2.2) é realizado. A diferença entre eles é que imagem FSeE é a face fonte neutra e não a imagem *ground truth*.

A métrica foi aplicada para os bancos IMPA e MUG. Os resultados do experimento descrito nesta seção são apresentados na seção 6.5.

4.2.3 Métrica subjetiva

A métrica subjetiva utilizada para avaliar as imagens sintetizadas está descrita no artigo Testa et al. (2018b) e é resumida nesta seção e na seção 6.1. Voluntários foram informados de que fotografias (imagens reais) e imagens geradas por computador (não reais) seriam avaliadas. Embora o mesmo conjunto de imagens tenha sido avaliado por todos os voluntários, cada participante recebeu um formulário contendo imagens em uma ordem aleatória diferente. Este procedimento visa evitar um viés relacionado à ordem de apresentação das imagens. Os voluntários deveriam categorizar uma determinada imagem como “real” ou “não real”, conforme ilustrado na figura 17. Cada imagem foi apresentada separadamente das demais para evitar comparações.

Figura 17 – Exemplo de questão exibida aos participantes.

2 How do you evaluate this image?

☐ Real

☐ Not real

Fonte: Adaptado de Testa et al. (2018b). Imagem originalmente sintetizadas a partir de imagens da base IMPA Faces(MENA-CHALCO; JUNIOR; VELHO, 2008).

As imagens utilizadas no formulário, além de sintéticas ou reais, incluíam uma das seis expressões faciais anteriormente especificadas. Assim, cada formulário continha as mesmas 36 imagens ($((3 \text{ sintéticas} + 3 \text{ reais}) \times 6 \text{ emoções})$). Os rostos apresentados nas imagens foram distribuídos igualmente entre os sexos masculino e feminino. Da mesma forma, nenhum rosto repetido foi exibido. As imagens apresentadas aos participantes foram obtidas ou sintetizadas a partir de imagens fornecidas em bancos de dados de expressões faciais descritos.

A avaliação do realismo considerou apenas dois níveis (“real” e “não real”), em vez de uma escala de realismo, porque não há evidências de que uma pessoa possa perceber graus de realismo (RADEMACHER, 2002). Além disso, o termo “não real” foi escolhido em vez de “sintético” ou “falso” porque é a negação direta de “real”, reduzindo assim o viés do significado desses termos (RADEMACHER, 2002).

Para este experimento, se a síntese foi eficaz para representar as imagens de forma realista, espera-se que o número de imagens sintéticas consideradas reais seja maior ou igual ao número de imagens reais consideradas reais. Para tanto, consideramos os “tipos de imagem”, sejam eles reais (fotografia) ou não (sintetizados a partir da abordagem proposta), para levantar as seguintes hipóteses:

H₀ a atribuição de realismo a uma imagem, por avaliadores voluntários, não depende do tipo de imagem (real ou sintética);

H₁ a atribuição de realismo a uma imagem, por avaliadores voluntários, depende do tipo de imagem (real ou sintética).

As hipóteses foram testadas usando o teste qui-quadrado de independência (COCHRAN, 1952).

Cento e cinquenta e cinco participantes responderam ao formulário (ver figura 17). Para obter uma avaliação adicional qualitativa, os dados foram coletados com dez voluntários de forma assistida, incluindo uma tarefa extra. Eles deveriam mostrar nas imagens quais fatores os levaram a considerar algumas imagens como “não reais”. As respostas a essa pergunta foram utilizadas para discutir os fatores que influenciam o realismo.

4.3 *Considerações finais*

Neste capítulo foram apresentados os motivos pelos quais cada técnica foi considerada em cada etapa da síntese de expressões faciais por “Deformação e Mapeamento” adotada por este trabalho, sendo elas: a seleção das imagens fontes, detecção facial, correspondência entre características faciais, alinhamento das faces, deformação e mapeamento. Os detalhes da implementação de tais etapas estão descritos no capítulo 5.

Além disso, foram delineados experimentos para avaliar os resultados, sendo um experimento subjetivo para avaliação do realismo e três experimentos objetivos avaliando as etapas envolvidas na síntese de expressões faciais: seleção das imagens fontes, síntese das expressões faciais considerando as diferentes abordagens para seleção das imagens fontes e preservação da identidade.

Por fim, também foram descritos os bancos de faces e expressões faciais utilizados nessas avaliações. Os resultados desses experimentos são apresentados no capítulo 6.

5 Síntese de expressões faciais por faces similares

Nesta seção são apresentados os métodos definidos para a síntese das expressões faciais, a métrica de avaliação dos resultados e as imagens que irão compor o banco de expressões faciais. O objetivo do método de síntese das expressões faciais é alterar uma fotografia fornecida (F_{dn}) de modo a gerar uma nova imagem contendo a expressão facial desejada (F_s), conforme descrito na seção 1. Para esse processo são necessárias outras duas imagens de outro indivíduo realizando a expressão facial desejada, denominadas faces fonte neutra (F_{fn}) e expressiva (F_{fe}). Ou seja, deseja-se construir uma função $f : I_f^E \times I_f^N \times I_d^N \rightarrow I_d^E$ em que, dadas uma imagem fonte I_f^E contendo uma expressão facial E , uma imagem fonte I_f^N contendo uma expressão neutra (N) do indivíduo de I_f^E e uma imagem destino I_d^N contendo uma expressão neutra de outro indivíduo, sintetiza-se uma imagem I_d^E que reconstitui a expressão E sobre a imagem destino, preservando as características faciais que representam sua identidade.

A figura 18 mostra as etapas necessárias para a síntese de acordo com a presente proposta. Essas etapas foram identificadas no capítulo 3 e a maneira como irão compor a abordagem é especificada nas próximas subseções. Todos os processamentos realizados nas imagens se referem à parte interna da delimitação do contorno da face (ver figura 24). Eles tratam das regiões da face como olhos, nariz, boca, testa, bochecha, etc; mas não regiões como cabelo, orelha, etc.

Figura 18 – Diagrama das etapas de sintetização pela presente proposta.

Fonte: Rafael Luiz Testa, 2018

5.1 Detecção facial

A detecção da face tem como objetivo determinar a localização da face na imagem. A detecção é realizada pela *MMOD* (Seção 2.3) utilizando as características extraídas por uma *CNN* (Seção 2.9), cuja arquitetura é apresentada na figura 19a.

Nesta etapa foi utilizado o mesmo modelo de Marčetić, Soldić e Ribarić (2017), treinado e implementado pela biblioteca DLIB¹ (KAZEMI; SULLIVAN, 2014). Ele foi treinado com um conjunto de 6.975 imagens faciais (KING, 2016) formado por imagens de diferentes bancos de faces, sendo eles: ImageNet (DENG et al., 2009), AFLW (KOESTINGER PAUL WOHLHART; BISCHOF, 2011), Pascal VOC (EVERINGHAM et al., 2015), *VGG face dataset* (PARKHI; VEDALDI; ZISSERMAN, 2015), WIDER (YANG et al., 2016), e *FaceScrub* (NG; WINKLER, 2014).

5.2 Detecção dos landmarks faciais

A identificação dos *landmarks* faciais foi utilizada em duas etapas: a seleção das imagens fonte (Seção 5.5) e a correspondência das características. Ao todo são 68 pontos distribuídos entre: sobrancelhas, olhos, nariz, boca e contorno da face; como exemplificado nas imagens da figura 20. A posição dos *landmarks* é detectada automaticamente por meio de *ensemble regression trees* (ver seção 2.4). Para esta etapa foram utilizados o modelo treinado da biblioteca DLIB² (KAZEMI; SULLIVAN, 2014), treinado com um conjunto de 300 faces do banco de imagens *I-BUG 300W* (SAGONAS et al., 2016). Ela possui um desempenho superior em relação a outras abordagens do estado da arte (WANG et al., 2017).

Os 68 *landmarks* identificados não se mostram suficientes para sintetizar imagens realistas, conforme identificado no experimento relatado na seção 6.1. Assim, 24 pontos foram adicionados nas bordas da imagem, como em Averbuch-Elor et al. (2017). Esses *landmarks* são importantes, por exemplo, quando alguma região da face, como o queixo, encolhe. Nesta situação, não deformar a área fora do rosto gera artefatos na imagem sintética.

Além disso, também foram adicionados 16 pontos nas seguintes regiões: ao lado dos olhos, bochechas, queixo, e testa. Esses *landmarks* foram importantes para a definição de uma máscara durante o mapeamento da iluminação (ver seção 5.7). Esses *landmarks* foram adicionados no ponto médio entre os dois *landmarks* que o cercam. Ao todo, foram adicionados 40 pontos, totalizando 108 *landmarks* como representado nas imagens da figura 20.

¹ Disponível em <http://dlib.net>

² Disponível em <http://dlib.net>

Figura 19 – Ilustração das arquiteturas das redes *CNN* para detecção facial e *ResNet* para descrição da face. “Conv” significa convolução, “TC” representa as camadas totalmente conectadas, o número anterior à “conv” representa o tamanho da janela do filtro, o posterior à “conv” quantidade de filtros, “/2” trata-se do *pooling* e as linhas tracejadas mostram os atalhos com aumento de dimensão.

(a) Arquitetura da *CNN* para reconhecimento facial. (b) Arquitetura da *ResNet* utilizada como descritor da face.

Fonte: Rafael Luiz Testa, 2018

Figura 20 – Identificação do *landmarks* faciais com exemplos. Os pontos em azul são os *landmarks* identificados, enquanto os pontos em vermelho foram adicionados com base na posição dos demais pontos.

Identificação dos *landmarks* faciais

Fonte: A fotografia original disponibilizada em *Pexel* por Daniel Xavier

5.3 Triangulação das faces

Além da identificação dos *landmarks* faciais, outro pré-processamento importante é o cálculo da triangulação. A triangulação foi calculada para apenas um conjunto de *landmarks* de uma face porque diferentes configurações faciais podem gerar diferentes triangulações. Esse fator é importante para correspondência entre os triângulos de duas faces, permitindo a transformação deles durante o alinhamento (Seção 5.4) e a deformação (Seção 5.6).

Esse processo foi inicialmente conduzido por meio da triangulação de *Delaunay*³ (Seção 2.6). Os triângulos gerados são exemplificados nas imagens da figura 21. A triangulação de *Delaunay* foi escolhida por dois motivos: os triângulos formados possuem ângulos parecidos e a triangulação resultante não depende da ordem em que os pontos

³ Implementação disponível em <http://opencv.org/>.

Figura 21 – Triangulação da face por *Delaunay* e ajustes manuais. Os triângulos destacados em vermelho na imagem da esquerda foram ajustados manualmente e conforme os triângulos em vermelho na direita.

Fonte: A fotografia original disponibilizada em *Pexel* por Daniel Xavier.

são processados (LI; ZHU; GOLD, 2004). Apesar disso, os triângulos da região da boca precisaram ser reajustados manualmente (ver figura 21) porque alguns triângulos cobriam tanto os lábios quanto a região interna da boca (dentes, língua, etc.). Entretanto, a região interna da boca é deformada a partir da face fonte (ver 5.6) e não da mesma imagem, o que pode gerar artefatos.

5.4 Alinhamento das faces

Essa etapa permite a normalização dos formatos da face, necessária para garantir um melhor resultado do mapeamento das características, descrito na atividade de “Mapeamento de iluminação” (Seção 5.7). Além disso, é uma etapa necessária para assegurar que o cálculo do novo formato utilizado na atividade de “Deformação da face” (Seção 5.6) seja consistente. O processo descrito nesta seção foi adaptado de Baggio et al. (2017) e suas etapas estão ilustradas na figura 22 e descritas a seguir.

Uma matriz de transformação da imagem é definida em termos de rotação e escala. A rotação é definida de modo que os centróides dos *landmarks* de ambos os olhos tenham a mesma posição horizontal, ou seja, a imagem é rotacionada de modo a compensar o ângulo encontrado. O ângulo entre os centróides é calculado conforme a equação 9, na

Figura 22 – Esquematisação das principais etapas do alinhamento para uma face.

Fonte: A fotografia original disponibilizada em *Pexel* por simpang.

qual são utilizadas as coordenadas x e y dos olhos esquerdo (O^e) e direito (O^d). A rotação é realizada em torno do ponto médio entre os centróides (Figura 22).

$$\Theta = \arctan\left(\frac{O_y^e - O_y^d}{O_x^e - O_x^d}\right) \quad (9)$$

Para a matriz de transformação também precisamos definir a escala. A escala garante que as distâncias entre os olhos sejam as mesmas. A escala está definida na equação 10, na qual $d_{desejada}$ é a distância desejada entre os centróides dos olhos definida pelo usuário, d_{atual} é a distância euclidiana entre os olhos atual e $l_{desejada}$ é a nova largura da imagem definida pelo usuário. d_{atual} é calculada pela distância euclidiana entre os centróides dos olhos, como mostra a equação 11. Além disso, é possível definir $d_{desejada}$ automaticamente escolhendo seu valor como a menor distância entre os olhos do conjunto de imagens processado.

$$escala = \left(\frac{d_{desejada}}{d_{atual}}\right) \cdot l_{desejada} \quad (10)$$

$$d_{atual} = \sqrt{(O_x^e - O_x^d)^2 + (O_y^e - O_y^d)^2} \quad (11)$$

Os parâmetros centro (ponto médio dos centróides), ângulo e escala são utilizados para calcular a matriz de transformação t ⁴, definida na equação 12, na qual $centro = \frac{O^e + O^d}{2}$ para x e para y (ver ponto médio na figura 22), $\alpha = escala \cdot \cos(\theta)$, e $\beta = escala \cdot \sin(\theta)$.

$$t = \begin{bmatrix} \alpha & \beta & (1 - \alpha) \cdot centro_x - \beta \cdot centro_y \\ -\beta & \alpha & \beta \cdot centro_x + (1 - \alpha) \cdot centro_y \end{bmatrix} \quad (12)$$

Na equação 12 é apresentada a matriz de rotação da biblioteca OpenCV. A matriz é alterada para adicionar o componente de translação da imagem. Ele centraliza a face vertical e horizontalmente na imagem. A translação é definida horizontalmente de modo que $centro_x$ se posicione no centro da imagem, e verticalmente ($centro_x$) na porção superior da imagem. A matriz atualizada (t') é definida na equação 13, na qual $l_{desejada} \cdot 0,5 - centro_x$ e $a_{desejada} \cdot 0,25 - centro_x$ são adicionados na terceira coluna. A transformação é aplicada por meio de uma transformação afin⁵ (Seção 2.7).

$$t' = t + \begin{bmatrix} 0 & 0 & l_{desejada} \cdot 0,5 - centro_x \\ 0 & 0 & a_{desejada} \cdot 0,25 - centro_x \end{bmatrix} \quad (13)$$

Por fim, durante a transformação afin a imagem também é recortada de modo a possuir uma largura ($l_{desejada}$) e altura ($a_{desejada}$) definida pelo usuário. Assim, elementos desnecessários são excluídos da imagem. Por exemplo, na figura 22 boa parte do fundo e a parte abaixo do pescoço são elementos removidos da imagem dispensáveis para síntese das expressões. Além disso, esse processo padroniza o tamanho das imagens, essencial para o mapeamento das características (Seção 5.7), pois é necessário que as imagens possuam as mesmas dimensões para obter-se a imagem razão (Seção 5.7).

5.5 Seleção de imagens fonte

O uso das imagens fonte pode ocorrer de diversas maneiras: a seleção manual, o uso de todas as imagens disponíveis para construção do modelo, ou a seleção automática de faces utilizando medidas de similaridade de modo a otimizar os resultados. Uma descrição

⁴ Implementação disponível em <http://opencv.org/>.

⁵ Implementação disponível em <http://opencv.org/>.

mais detalhada sobre essas abordagens é encontrada na seção 3.1. Para esta proposta foi realizada uma seleção automática das imagens fonte.

Ao escolher a seleção automática, a abordagem proposta permite a automatização da escolha das imagens fonte. A maior parte das técnicas de “Deformação e Mapeamento” encontradas na literatura fazem a seleção dessas imagens de forma manual (ver seção 3.1). Assim, incluir a escolha das faces fonte na abordagem permite uma automatização da síntese das expressões faciais, constituindo um diferencial do presente projeto.

Existem diferenças nos componentes faciais para cada pessoa, e isso resulta em diferenças nas expressões faciais. A fim de sintetizar imagens realistas, é preciso considerar as especificidades dos componentes faciais para cada indivíduo (XIE et al., 2015; LI et al., 2014). Para considerar esta questão, o presente projeto se propõe a fazer uma seleção das imagens fonte de modo que a face selecionada seja aquela do banco de imagens com maior semelhança com a face destino.

Os bancos de expressões faciais possuem imagens diferentes da mesma pessoa, sendo uma imagem neutra (sem expressão que represente emoções) e uma imagem que expresse cada emoção universal (satisfação, tristeza, raiva, aversão, medo e surpresa), conforme especificado na seção 4.2.1. Para a seleção das imagens fonte, a ideia é encontrar, dentre as imagens de expressão facial neutra presentes nesta base de imagens, aquela que apresenta as características faciais mais parecidas com a imagem da F_{dn} .

A seleção da F_{fn} é composta pelas etapas: i) extração das características faciais (descriptor) das imagens fontes, ii) criação de um índice a partir dessas características, iii) extração das características da F_{dn} , vi) comparação das características extraídas para F_{dn} com cada F_{fn} presente no banco de imagens, e v) seleção da face expressiva correspondente; conforme ilustra a figura 23.

A primeira etapa, extração de características, tem como entrada uma imagem facial e retorna um vetor com 128 características representando a face. Esse vetor é o resultado da *ResNet* (HE et al., 2016), uma *CNN* residual profunda similar à *ImageNet* proposta por He et al. (2016). A *ResNet* utilizada neste trabalho é estruturada em 29 camadas em contraste com a *ImageNet* que possui 34 camadas e metade da quantidade de filtros nas convoluções. Sua arquitetura é apresentada na figura 19b. A seção 2.9 apresenta mais detalhes sobre as camadas.

A rede foi treinada para extrair características faciais pensando em aplicações de reconhecimento facial, ou seja, ela é treinada com diversas imagens faciais de um mesmo

Figura 23 – Esquematização das etapas empregadas para encontrar uma imagem facial similar.

Fonte: As fotografias utilizadas nas imagens foram obtidas de *Pexel* disponibilizadas por Simon Roben, Ezekixl Akinnewu, e Photo by Min An from Pexels.

indivíduo, considerando vários indivíduos. Este modelo foi treinado e implementado pela biblioteca DLIB⁶ (KAZEMI; SULLIVAN, 2014). Ele foi treinado a partir de imagens dos bancos *VGG face dataset* (PARKHI; VEDALDI; ZISSERMAN, 2015), *FaceScrub* (NG; WINKLER, 2014) e imagens retiradas da internet pelo autor da biblioteca; possuindo aproximadamente três milhões de faces. Mais detalhes sobre o banco podem ser encontrados em King (2017).

Na segunda etapa, os vetores formados pelas características extraídas de cada F_{fn} compõem um índice de busca. Esse índice é armazenado em um arquivo no formato HDF5 (*Hierarchical Data Format*) (FOLK et al., 2011) para uso futuro.

Na terceira etapa, as características da F_{dn} são extraídas. Para a quarta etapa, o vetor composto por essas características é comparado com os vetores armazenados no índice. Para isso, é calculada a distância euclidiana entre o vetor da F_{dn} e os vetores do

⁶ Disponível em <http://dlib.net>

índice, conforme cálculo de similaridade apresentado na seção 2.5. A imagem facial neutra que possuir a menor distância em relação à F_{dn} é selecionada como F_{fn} .

Por fim, na quinta etapa, é selecionada como face fonte expressiva a F_{fe} equivalente à F_{fn} encontrada. Exemplos de faces similares expressivas selecionadas são apresentados na figuras 37 e 38.

5.6 Deformação da face

Nesta atividade os componentes da face destino neutra são transformados de acordo com as mudanças nos posicionamentos de seus *landmarks*. A mesma técnica de transformação geométrica utilizada para o alinhamento das faces (Seção 5.4) foi utilizada para a deformação da face.

O processo de deformação utiliza os *landmarks* faciais para a estimativa do novo formato de cada componente da face destino, considerando as diferenças entre a face fonte neutra e a face fonte expressiva. Para isso, calcula-se a diferença entre a posição do *landmark* na face fonte neutra e o correspondente na face fonte destino. Esse conjunto de diferenças nas posições dos pontos é, então, somado à expressão destino neutra. A etapa descrita nesse parágrafo corresponde à primeira etapa apresentada na figura 24 (Cálculo do conjunto expressivo de *landmarks*).

Na segunda etapa, os *pixels* de cada triângulo da face destino neutra são mapeados para o triângulo equivalente resultante da primeira etapa por meio de transformações afins, conforme apresentado na etapa dois da figura 24. Não é possível deformar as regiões que não estão sendo expostas na imagem, no caso do presente trabalho a região interna da boca. Para suprir esse componente ocluso na face destino, a mesma região da face fonte expressiva é utilizada na composição da imagem sintetizada. Para isso, é conduzido o mesmo processo de transformação afim dos demais componentes da face para a região interna da boca, porém agora temos a imagem fonte expressiva como base para a deformação, conforme apresentado na etapa três da figura 24.

No experimento apresentado na seção 6.1 foi encontrado que a diferença entre a parte interna da boca e os lábios pode gerar artefatos, como apresentado na imagem “Destino deformada” da etapa quatro da figura 24. Este problema foi solucionado com uma abordagem similar à Averbuch-Elor et al. (2017). Diferentemente de Averbuch-Elor

Figura 24 – Esquematisação da deformação da face neutra para expressiva. Ilustrações e exemplos esquematizam as etapas de cálculo dos *landmarks* expressivos, deformação da face, deformação da parte oclusa e mesclagem da parte oclusa.

Fonte: Imagens sintetizadas a partir de imagens da base IMPA Faces(MENA-CHALCO; JUNIOR; VELHO, 2008) e de imagens de Daniel Xavier e de Moose Photos *Pexels*.

et al. (2017), no presente trabalho o filtro é aplicado apenas nas bordas dos lábios para fazer a mesclagem da parte interna da boca com os lábios. Primeiramente, seleciona-se os *landmarks* dos lábios superiores e inferiores. Ou seja, de todos os *landmarks* separa-se apenas o subconjunto que contém os pontos de ambos os lábios. Em seguida, é criada uma

máscara da região de interesse desenhando os contornos ⁷. Por fim, um filtro de desfoque gaussiano (*gaussian blur*) é aplicado na região da máscara para a mesclagem das regiões.

5.7 Mapeamento da iluminação

Após a deformação dos componentes da face, é necessário mapear as informações de iluminação para ajustar a aparência da expressão facial sintetizada. O mapeamento possibilita a síntese de aspectos mais finos, como os enrugamentos. O método mais popular para o mapeamento é a transferência das mudanças de iluminação entre imagens faciais, a ERI (Seção 3.2). Nele, as mudanças de iluminação entre duas imagens são transferidas para uma terceira imagem com o objetivo de reproduzir essas mudanças na terceira imagem, conforme especificado no conceito de imagens razão (Seção 2.8).

A razão das imagens fonte captura as diferenças de iluminação entre as duas imagens fonte, para garantir melhores resultados. Parte-se do princípio que essas imagens foram obtidas nas mesmas condições de iluminação. Assim, a razão entre elas se refere apenas às mudanças relacionadas à expressão facial. Tal requisito pode ser alcançado com, por exemplo, dois *frames* de uma filmagem. Esse requisito é atendido pelas imagens presentes no banco de expressões faciais. Então, a imagem razão é multiplicada pela face destino deformada para obter a imagem final, que contém a expressão facial desejada. Esse processo é formulado nas equações 3 e 15; e ilustrado na figura 25.

Primeiramente, as imagens são convertidas do formato *RGB* para *YCbCr*. Então, apenas o componente de luminância (*Y*) é utilizado para o cálculo da imagem razão. O uso de apenas este canal pode assegurar melhor resultados (LIU; SHAN; ZHANG, 2001). A razão é, então, calculada entre as luminâncias da face fonte expressiva (Y_{fe}) e da face fonte neutra (Y_{fn}), conforme formulado na equação 14, em que x e y correspondem às coordenadas dos *pixels* nas imagens.

$$R = \frac{Y_{fe}(x, y)}{Y_{fn}(x, y)} \quad (14)$$

Na figura 25 é possível perceber que mesmo usando imagens fontes controladas, é gerada uma imagem razão com elementos que não correspondem apenas às alterações faciais relativas à expressão facial. Por isso, foi necessário realizar uma filtragem dos

⁷ Implementação disponível em <http://opencv.org/> para os pontos do lábio superior e para os pontos do lábio inferior

Figura 25 – Esquematização das etapas empregadas para mapeamento das características, envolvendo a imagem razão, melhorias nas imagem razão e a síntese da imagem expressiva. A imagem razão foi invertida para uma melhor visualização, ou seja, os *pixels* pretos equivalem a um e os *pixels* brancos a zero.

Fonte: Imagens originais retiradas ou sintetizadas a partir de imagens da base IMPA Faces(MENACHALCO; JUNIOR; VELHO, 2008).

elementos que não se referiam às mudanças de uma expressão facial, como a mudança de posição do cabelo. Para isso, em um processo similar à Averbuch-Elor et al. (2017), um filtro de desfoque gaussiano é aplicado à imagem razão. A imagem desfocada é utilizada como índice para remover os *pixels* de maior e menor intensidades da imagem razão original. Nesta implementação, restam apenas os *pixels* maiores que 0.05 e menores que 0.95.

A imagem razão sem os extremos ainda possui elementos indesejados, como, por exemplo, a região do cabelo (ver figura 5.7). Estas regiões não desejadas foram removidas aplicando uma máscara para selecionar apenas as que possuem enrugamentos. Essa máscara foi construída preenchendo os triângulos nas regiões em que sabe-se previamente que há enrugamentos (EKMAN; FRIESEN; HAGER, 2002). Os procedimentos apresentados nessa seção foram definidos conforme Impett, Robinson e Baltrusaitis (2014), Averbuch-Elor et al. (2017). No entanto, a técnica desenvolvida para o mascaramento é um diferencial implementado que proporcionou melhores resultados nos experimentos conduzidos.

A luminância da face destino deformada (Y_{dd}) é multiplicada pela imagem razão (R), gerando a luminância da imagem da face sintética (Y_s), conforme formulado na equação 15. Então, a luminância gerada mais Cr (diferença de vermelho) Cb (diferença de azul) originais são convertidas novamente para RGB para gerar a F_s .

$$Y_s(x, y) = R \cdot Y_{dd}(x, y) \quad (15)$$

5.8 Considerações finais

O presente capítulo detalhou a definição de cada uma das etapas desenvolvidas no presente trabalho para síntese das expressões faciais, assim como suas relações com os conceitos relacionados (capítulo 2). Este capítulo também apresentou aspectos importantes para implementação de tais etapas como bibliotecas de código aberto, bancos de imagens e modelos utilizados.

No próximo capítulo são apresentadas e analisadas as imagens sintetizadas a partir das especificações do presente capítulo. Também são apresentados e discutidos os resultados referentes aos experimentos delineados no capítulo 4.

6 Resultados e discussões

Os resultados aqui apresentados mostram exemplos e analisam as imagens sintetizadas com base no método proposto (Capítulo 5), além de expor os resultados obtidos a partir das métricas descritas nas seções 4.2.2 e 4.2.3. Na seção 6.1 são apresentados os resultados de um experimento inicial para levantar os fatores que levam os indivíduos a perceberem as imagens sintetizadas como não reais. Os resultados desse experimento possibilitaram melhorar aspectos importantes no método de síntese das expressões faciais. As alterações nos métodos para incluir essas melhorias estão descritas no capítulo 5, e as imagens sintetizadas após essas alterações são apresentadas na seção 6.2. Os resultados referentes às imagens finais foram avaliadas pelas métricas objetivas: avaliação da busca por face similar (Seção 4.2.2), avaliação da síntese com face similar (Seção 4.2.2), avaliação da preservação da identidade (Seção 4.2.2); e estão descritas, respectivamente, nas seções 6.3, 6.4, e 6.5.

6.1 *Avaliação da percepção de realismo*

Nesta avaliação, são apresentadas as imagens sintéticas sem as alterações de melhoria, conforme descritas no capítulo 5. A seção 6.1.1 apresenta uma análise da atribuição de realismo às imagens. A seção 6.1.2 discute aspectos do realismo relevantes para a síntese da expressão facial. Por fim, a seção 6.1.3 aborda o realismo de cada expressão facial. Este experimento foi delineado na seção 4.2.3.

6.1.1 Análise estatística

Os tempos de resposta para imagens reais e sintéticas foram similares. Imagens sintéticas foram respondidas em média em 10,0 segundos (DP=2,4), enquanto as imagens reais tiveram uma média de 9,8 segundos (DP=4,5).

Como discutido em 4.2.3, um teste de independência qui-quadrado foi realizado para examinar se existe relação entre realismo e tipo de imagem. O teste de independência entre essas variáveis foi não significativo, $\chi^2(1, N = 115) = 1229,45$, $p > 0,05$, indicando que o realismo depende do tipo de imagem. A figura 26 apresenta a porcentagem média

Figura 26 – Porcentagem média de atribuições como “real” e sintética (“não real”) para cada uma das imagens.

(a) Imagens reais

(b) Imagens sintéticas

Fonte: Adaptado de Testa et al. (2018b).

de atribuição de realismo para cada uma das imagens avaliadas (Figura 26a). Esta figura apresenta dois gráficos de atribuição de realismo: um para as imagens reais (Figura 26a) e outro para as imagens sintéticas (Figura 26b). As atribuições de realismo para imagens reais (Figura 26a), em geral, foram maiores que para as sintéticas (Figura 26b), sugerindo que as imagens sintetizadas precisavam de melhorias para serem percebidas como reais. Isso não significa que tais imagens não poderiam ser usadas nas aplicações reportadas no capítulo 1, mas elas precisam de melhorias para que uma pessoa não possa distinguir entre uma imagem sintética e uma imagem real.

Na tabela 4, é possível observar que alguns voluntários que participaram da avaliação consideraram imagens sintéticas como reais (30%). Alguns participantes também consideraram imagens reais como sintéticas (23%). O gráfico da figura 26 mostra que as três imagens reais com as atribuições mais altas como “não reais” têm porcentagens similares às três imagens sintéticas com a maior atribuição de realismo. A seção 6.1.2 discute e exemplifica os fatores que podem ter influenciado a percepção do realismo para essas imagens.

Um estudo relacionado (AVERBUCH-ELOR et al., 2017) conduziu uma avaliação na qual a atribuição de realismo é avaliada em vídeos sintéticos ou reais. Diferentemente da presente avaliação, o trabalho considera cinco níveis diferentes para atribuição de realismo: muito provavelmente falso (1), provavelmente falso (2), igualmente real ou falso (3), provavelmente real (4) e muito provavelmente real (5). Para vídeos sintéticos, os

Tabela 4 – Matriz de confusão das respostas. As linhas referem-se ao tipo de imagem (sintéticas ou reais) e a todas as imagens independentemente do tipo, enquanto as colunas referem-se às respostas (sintéticas ou reais) e ao total de respostas.

	Respostas		
	Sintéticas	Reais	Total de respostas
Sintéticas	70% (1945)	30% (845)	100% (2790)
Reais	23% (638)	77% (2152)	100% (2790)
Todas imagens	46% (2583)	54% (2997)	100% (5580)

autores encontraram que 34% dos entrevistados atribuíram os níveis um e dois (tendendo a falsos), 46% atribuíram os níveis quatro e cinco (tendendo a real) e 20% estavam indecisos.

Essa abordagem comparada soluciona alguns problemas relacionados ao realismo das imagens sintéticas, como bordas irregulares e rugas, conforme discutido na seção 6.1.2. Depois deste experimento essas soluções também foram adotadas pela metodologia proposta (Capítulo 5). O experimento do trabalho comparado (AVERBUCH-ELOR et al., 2017) e deste experimento também têm diferenças na maneira de coletar dados de atribuição de realismo (cinco níveis de realismo contra dois níveis do presente trabalho) e no uso da palavra falsa no trabalho citado em vez do termo “não real” pela presente abordagem. Além disso, os estudos também tiveram um número diferente de voluntários: 30 em Averbuch-Elor et al. (2017) versus 155 no presente experimento.

Essas diferenças nas avaliações dificultam uma comparação direta entre os resultados. Apesar das diferenças apontadas, resultados semelhantes foram obtidos na atribuição de realismo para este experimento e nos níveis quatro e cinco para o estudo comparado (AVERBUCH-ELOR et al., 2017). Para essa comparação, foram considerados todos os quatro vídeos de (AVERBUCH-ELOR et al., 2017) comparados com as quatro imagens que obtiveram os melhores resultados de atribuição de realismo em relação às mesmas expressões (ver figura 26). Os resultados de ambos os trabalhos são apresentados na 5. As avaliações subjetivas de ambos os trabalhos apresentaram resultados similares. Os resultados obtidos pelo presente experimento foram um pouco abaixo em relação ao estudo comparado (AVERBUCH-ELOR et al., 2017). Entretanto as imagens avaliadas nesse experimento não correspondem às imagens sintéticas finais pela abordagem proposta, e sim às imagens sintéticas antes das melhorias propostas no capítulo 5.

A escolha de comparar os quatro melhores resultados se deu porque o estudo relacionado (AVERBUCH-ELOR et al., 2017) não especifica em quais imagens/vídeos a

Tabela 5 – Comparação das porcentagens de atribuição de realismo das quatro imagens com melhores resultados obtidos (ver figura 26) em relação à avaliação dos quatro vídeos de Averbuch-Elor et al. (2017).

	Resultados obtidos			
	Satisfação	Medo	Surpresa	Raiva
Resultados obtidos	52%	42%	37%	32%
Averbuch-Elor et al. (2017)	57%	49%	36%	41%

avaliação foi realizada. Assim, é difícil estabelecer uma comparação precisa dos resultados. Também é importante ressaltar que o estudo comparado (AVERBUCH-ELOR et al., 2017) avalia vídeos e não imagens, que são mais complexas devido às características relacionadas ao processo de animação.

6.1.2 Fatores que influenciam a percepção do realismo

Os fatores que influenciaram a percepção de realismo foram: “diferenças entre regiões subsequentes” e “aspectos não naturais em regiões da face”. Essas categorias permitem uma análise conjunta de fatores semelhantes, sendo a seguir ilustrados com imagens sintéticas e reais.

Diferenças entre regiões subsequentes

O primeiro artefato envolve bordas irregulares proeminentes. A irregularidade pode aparecer como resultado da descontinuidade entre duas regiões que sofreram o processo de deformação. Os participantes apontaram para bordas irregulares dentro da boca (Figura 27a) e no queixo (Figura 27c), em contraste com a figura 27b que não possui tais artefatos. Esse artefato influenciou negativamente a percepção do realismo. Ele pode ter duas causas diferentes dependendo da região em que está localizado. Na parte interna da boca, o artefato é gerado pela diferença entre F_{fe} e F_{dn} , uma vez que a região interna da boca é deformada a partir de F_{fe} (ver etapa três da figura 24). Aplicar filtros na borda dos lábios pode diminuir esse problema. Para a região do queixo, o artefato é gerado porque apenas a região interna da face é deformada. Assim, para algumas imagens, o artefato pode aparecer porque a região do queixo é encolhida, enquanto a parte abaixo não é alterada. A solução para esse problema seria aplicar transformações afins nas regiões da imagem que estão

Figura 27 – Exemplos de imagens com e sem bordas irregulares na região da boca e em regiões com descontinuidade.

Fonte: Adaptado de Testa et al. (2018b). Imagens originalmente retiradas ou sintetizadas a partir de imagens da base IMPA Faces(MENA-CHALCO; JUNIOR; VELHO, 2008).

fora da face. Isso pode ser feito incluindo *landmarks* e triângulos no pescoço ou na borda da imagem.

Outro artefato desta categoria refere-se a uma notável diferença de cor entre regiões subsequentes da imagem. Curiosamente, essa diferença foi um fator que contribuiu para que os voluntários atribuísem a ausência de realismo tanto em imagens sintéticas e reais. Um exemplo é apresentado na figura 28a, que mostra sombreamento na região do pescoço. Outro exemplo é a cor mais escura dentro da boca (figura 28c e figura 28d). Voluntários afirmaram que esses aspectos evidenciaram que a imagem havia sido alterada digitalmente e que a região não havia sido adequadamente tratada. Nas imagens reais, essa diferença provavelmente ocorreu por causa das condições de iluminação durante a aquisição da imagem. Essas imagens reais foram usadas como fontes para sintetizar outras imagens; assim, o mesmo problema foi percebido nas duas imagens (sintética e real). Contudo, a diferença de cor entre F_{sn} e F_{te} também contribui para aumentar a diferença.

Aspectos não naturais em regiões da face

Outro tipo de artefato relatado nas imagens são aspectos percebidos como não naturais em imagens faciais. Um exemplo é um tamanho de pupila muito grande, como indicado em imagens sintéticas (Figura 29a) e reais (Figura 29c). Seyama e Nagayama (2007) avaliou o efeito de *Uncanny valley* em imagens reais e sintéticas, também encontrou uma menor atribuição de realismo às imagens que possuem componentes faciais maiores. *Uncanny valley*, também conhecido com vale da estranheza, é um efeito que acontece em

Figura 28 – Exemplos de imagens com diferença de cor. Imagens retiradas ou sintetizadas a partir de imagens de IMPA Faces.

Fonte: Adaptado de Testa et al. (2018b). Imagens originalmente retiradas ou sintetizadas a partir de imagens da base IMPA Faces(MENA-CHALCO; JUNIOR; VELHO, 2008).

faces virtuais muito parecidas com faces reais. Os pequenos aspectos não naturais levam as pessoas a uma sensação de estranheza. A desproporção relatada neste experimento pode ser causada por dois fatores: (i) a deformação de toda a região ocular (e não apenas a pupila) para sintetizar a imagem; e (ii) o fato das expressões faciais nos bancos de imagens serem encenadas (sem expressões espontâneas), o que pode causar uma abertura maior ou menor que o normal para o olho. Na abordagem proposta, a região da pupila não é tratada de forma independente do olho. A alteração desproporcional da pupila pode ser melhorada com a identificação e deformação desta região separadamente do olho. O segundo problema pode ser resolvido com a construção de um banco imagens de expressões faciais espontâneo. As expressões faciais poderiam parecer mais congruentes, portanto, também poderiam auxiliar na minimização desse artefato; entretanto, questões adicionais, como o alinhamento facial e a dificuldade de obter imagens frontais de expressões faciais espontâneas teriam que ser abordadas.

Outro problema relatado pelos voluntários está na “complexidade” da expressão facial. Alguns participantes apontaram que as expressões faciais com rugas eram reais porque seriam mais difíceis de gerar pelo computador. Este aspecto pode ser observado ao comparar as figuras 30a e 30b. De fato, a abordagem até o momento não havia sintetizado enrugamentos. Para conseguir isso, técnicas como *ERI* (LIU; SHAN; ZHANG, 2001) foram

Figura 29 – Exemplos de componentes faciais com tamanhos não naturais.

Fonte: Adaptado de Testa et al. (2018b). Imagens originalmente retiradas ou sintetizadas a partir de imagens da base IMPA Faces(MENA-CHALCO; JUNIOR; VELHO, 2008).

Figura 30 – Exemplos de expressões faciais com e sem enrugamentos.

Fonte: Adaptado de Testa et al. (2018b). Imagens originalmente retiradas ou sintetizadas a partir de imagens da base IMPA Faces(MENA-CHALCO; JUNIOR; VELHO, 2008).

posteriormente adicionadas para sintetizar expressões faciais de forma mais convincente (Capítulo 5).

Os aspectos apresentados nesta seção indicam possíveis melhorias na abordagem proposta. Artefatos como bordas serrilhadas, diferenças de cores e características faciais não naturais refletem aspectos do realismo que devem ser observados ao propor abordagens de síntese de expressões faciais nas quais o realismo é um pré-requisito importante. No entanto, alguns desses artefatos exigem melhorias não apenas na síntese de uma nova imagem, mas também nos bancos de expressões faciais.

6.1.3 Análise das expressões faciais

A atribuição de realismo também varia de acordo com a expressão facial. Por exemplo, a figura 31b mostra a expressão facial de satisfação que obteve a maior atribuição de realismo por voluntários entre as imagens sintéticas. Entre as imagens reais (Figura 31a), a expressão facial de aversão (Figura 32a) alcançou a maior atribuição de realismo.

Dyck et al. (2008) avaliou o reconhecimento de expressões faciais em imagens reais e sintéticas. A expressão facial de satisfação alcançou a maior taxa de reconhecimento, e a menor foi alcançada para imagens com a expressão de aversão. A expressão facial

Figura 31 – Porcentagem de respostas como “real” e sintéticas (“não real”) para cada expressão facial.

(a) Imagens reais

(b) Imagens sintéticas

Fonte: Adaptado de Testa et al. (2018b).

Figura 32 – Exemplos de expressões faciais.

(a) Imagem real de aversão.

(b) Imagem sintética de satisfação.

(c) Imagem sintética de tristeza.

Fonte: Adaptado de Testa et al. (2018b). Imagens originalmente retiradas ou sintetizadas a partir de imagens da base IMPA Faces(MENA-CHALCO; JUNIOR; VELHO, 2008).

de satisfação também foi a expressão facial sintética mais classificada como real neste trabalho e em um trabalho relacionado (AVERBUCH-ELOR et al., 2017). A expressão facial de aversão foi a expressão com maior realismo entre os rostos reais. Uma possível explicação para isso é que expressões faciais “complexas” tendem a ser consideradas mais reais (ver seção 6.1.2). A expressão facial de aversão (Figura 32a) tem rugas em várias partes do rosto, enquanto a satisfação tem pequenas rugas apenas nos cantos dos olhos.

Para as imagens sintéticas (Figura 31b) as duas expressões faciais que obtiveram a maior atribuição de realismo foram satisfação e tristeza. Para estas expressões, a região da boca é alterada de forma bastante significativa, especialmente no canto dos lábios. A maior atribuição de realismo a essas expressões faciais pode ser explicada porque a síntese foi

eficaz em alterar essa região (boca); exemplos dessas expressões faciais são apresentados nas imagens das figuras 32b e 32c. Isso pode ter ocorrido porque a boca tem um número maior de *landmarks* e triângulos. O componente facial da boca possui 20 pontos, enquanto os outros componentes faciais apresentam quantidade menor de pontos: nariz (9 pontos de referência), olho (6 pontos de referência) e sobrancelha (6 pontos de referência).

6.1.4 Melhorias após experimento

O experimento apresentado neste capítulo ajudou a entender quais aspectos relacionados ao realismo poderiam ser melhorados. Dos aspectos elucidados na seção 6.1.2, foram acrescentados *landmarks* nas bordas da imagem (Seção 5.2, mais especificamente na figura 20) que resultaram na diminuição das bordas irregulares na região interna da boca, como exemplificado na figura 33a quando comparada à figura 33b. Também foi realizada uma mesclagem da parte interna da boca (Seção 5.6, mais especificamente na etapa 4 da figura 24) que resultou na diminuição das bordas irregulares desta região, como exemplifica a figura 33d em contraste com a imagem da figura 33c. O último aspecto que recebeu melhorias foram os enrugamentos.

Figura 33 – Exemplos de imagens anteriores e posteriores as melhorias em relação às bordas irregulares proeminentes.

(a) Imagem do experimento com bordas irregulares proeminentes na região do queixo.

(b) Imagem final sem bordas irregulares proeminentes na região do queixo.

(c) Imagem do experimento com bordas irregulares proeminentes na região interna da boca.

(d) Imagem final sem bordas irregulares proeminentes na região interna da boca.

Fonte: Imagens sintetizadas a partir de imagens da base IMPA Faces(MENA-CHALCO; JUNIOR; VELHO, 2008).

A abordagem de mapeamento da iluminação (seção 5.7), em alguns casos, foi capaz de sintetizar os enrugamentos que melhoram a percepção de realismo, conforme ilustrado na figura 34. A imagem da figura 34a é a imagem base para síntese das imagens das figuras 34b, 34c e 34d. A imagem da figura 34c foi sintetizada antes das melhorias descritas. A imagem da figura 34b foi sintetizada após as melhorias, porém sem os enrugamentos, e, por fim, a imagem da figura 34d é a imagem final depois das melhorias e com enrugamentos. É possível notar que a imagem da figura 34d apresenta enrugamentos mais acentuados na região abaixo do olho e das bochechas em contrastes com as figuras 34a, 34b e 34c. Além dessa figura, a imagem da figura 33b apresenta um enrugamento mais acentuado da região do queixo em relação à imagem da figura 33a. O mesmo acontece com a imagem da figura 33c em relação à imagem da figura 33d.

Figura 34 – Exemplos de imagens anteriores e posteriores a adição dos enrugamentos nas imagens sintéticas.

(a) F_{dn} do banco de imagens.

(b) Face destino deformada final (antes da adição dos enrugamentos).

(c) F_s do experimento.

(d) F_s final.

Fonte: Imagens sintetizadas a partir de imagens da base IMPA Faces(MENA-CHALCO; JUNIOR; VELHO, 2008).

Apesar de os enrugamentos terem recebidos melhorias, a abordagem não foi capaz de produzir os enrugamentos necessários para algumas imagens, como no caso da região

da testa da figura 30b. Assim, para esta imagem, a região permanece inalterada. Duas soluções ajudariam a melhorar esse aspecto: uma abordagem mais robusta para síntese dos enrugamentos e imagens fontes expressivas que possuam enrugamentos mais expressivos nessa região (ver seção 6.1.2).

Mais melhorias seriam necessárias para resolver os demais aspectos relacionados ao realismo levantados na seção 6.1.2. Não foram implementadas melhorias relacionadas à diferença de cor na região do pescoço (Figura 28a), diferença de cor na região da boca (Figura 28c), e componentes naturais com tamanhos artificiais (Figura 29a). Assim, tais aspectos permanecem iguais nas imagens sintéticas finais (Seção 6.2).

Ainda que não tenha sido proposta uma solução direta para a falta de expressividade de algumas imagens sintéticas, a seleção das imagens fontes conseguiu melhorar esse aspecto ao selecionar a imagem mais parecida (ver seção 5.5). A figura 33d mostra uma imagem exemplificando o aumento da expressividade de uma imagem sintética, nessa imagem a região da boca possui uma abertura maior que a imagem anterior na figura 33c. Assim, a imagem gerada fica mais parecida com uma expressão facial da emoção de raiva.

6.1.5 Demais melhorias

Esta seção descreve os aperfeiçoamentos da abordagem que não foram encontrados em outros trabalhos, por isso, se constituem como diferenciais. A seleção das imagens fontes foi capaz de melhorar a expressividade (Seção 6.1.4), preservação da identidade (Seção 6.5) e reduzir a quantidade de artefatos gerados (Seção 6.4). Além dessa contribuição, outras etapas da síntese foram melhoradas de modo a sintetizar imagens mais realistas, sendo elas a triangulação manual dos landmarks faciais, o alinhamento das imagens fontes e o mascaramento no processo de mapeamento da iluminação. As contribuições relativas a estas etapas são descritas a seguir.

O principal problema da triangulação de *Delaunay* era que, em alguns casos, ela gerava artefatos na região interna da boca. Isso não acontecia em todas as imagens, pois os triângulos variavam de acordo com a abertura da boca. Tais artefatos eram resultados de triângulos que abrangiam a região dos lábios juntamente com a parte interna da boca. A imagem da figura 35a ilustra um exemplo desse tipo de artefato, e a imagem da figura 35b ilustra o resultado após a triangulação manual.

Figura 35 – Exemplos de imagens anteriores e posteriores a triangulação manual.

(a) F_s antes dos ajustes dos triângulos. (b) F_s após os ajustes dos triângulos.

Fonte: Imagens sintetizadas a partir de imagens da base IMPA Faces(MENA-CHALCO; JUNIOR; VELHO, 2008).

Apesar de o alinhamento das faces gerar melhorias tanto na deformação quanto na aplicação do mapeamento da iluminação (ver seção 5.4), muitos trabalhos não se preocupam em implementar ou descrever uma técnica relacionada. O fato de alguns bancos de imagens já disponibilizarem as imagens previamente alinhadas reduz a necessidade da aplicação de técnicas relacionadas. Contudo, uma aplicação real da síntese de expressões faciais exigiria o alinhamento das imagens, pois a imagem de entrada (F_{dn}) não estará na mesma condição.

Em relação a mapeamento da iluminação foi adicionado uma máscara para selecionar as regiões a serem mapeadas (Seção 5.7, mais especificamente a etapa dois da figura 25). Essas regiões foram selecionadas, pois são as regiões em que os enrugamentos aparecem (EKMAN; FRIESEN; HAGER, 2002). A imagem da figura 36b foi sintetizada a partir da imagem deformada da figura 34b com os enrugamentos produzidos pela imagem razão ilustrada na figura 36a. Já a figura 36d foi sintetizada a partir da imagem deformada da figura 34b com os enrugamentos produzidos pela imagem razão ilustrada na figura 36c. Nota-se que a imagem da figura 36d possui menos artefatos em relação à imagem da figura 36b, especialmente na região das orelhas e cabelo.

6.2 Imagens sintetizadas

Nesta seção são apresentadas as imagens sintetizadas após as correções descritas no capítulo 5 referente às melhorias identificadas na avaliação subjetiva (Seção 6.1). As figuras 37 e 38 mostram exemplos de expressões faciais sintéticas a partir de imagens selecionadas por diferentes abordagens, sendo elas: características faciais (Seção 5.5), histograma de cor

Figura 36 – Exemplos de imagens anteriores e posteriores ao mascaramento. Os enrugamentos das figuras a e b foram sintetizados a partir da face deformada apresentada na imagem da figura 34b.

(a) Imagem razão sem mascaramento.

(b) F_s com enrugamento sintetizado sem mascaramento.

(c) Imagem razão com mascaramento.

(d) F_s com enrugamento sintetizado com mascaramento.

Fonte: Imagens sintetizadas a partir de imagens da base IMPA Faces(MENA-CHALCO; JUNIOR; VELHO, 2008).

(Seção 4.2.2), e aleatória (Seção 4.2.2). Para estas figuras todas as imagens sintetizadas possuíam a mesma face fonte como base.

As figuras 37e e 37f possuem uma menor abertura da boca, configurando uma expressividade menor do que a desejada (*ground truth*), porém um pouco mais próxima da face fonte expressiva (encontrada pela similaridade por características faciais). A região do nariz na figura 37f é levemente deformada para o lado esquerdo e na figura 37e para baixo. Apenas as figuras 37d e 37f tiveram o tamanho do olho reduzido como na imagem *ground truth*. Nenhuma das imagens sintéticas possui, na região interna da boca, iluminação similar à imagem *ground truth*, visto que este aspecto não é tratado pela abordagem proposta. Ou seja, para melhores resultados a diferença de iluminação da face fonte expressiva para face destino deveria ser considerada. Apesar do mapeamento da iluminação (Seção

Figura 37 – Exemplo de imagens sintéticas de satisfação para imagens selecionadas pelas abordagens características faciais, histograma de cor, e aleatória. À direita são apresentadas a face destino neutra e *ground truth*.

Fonte: Imagens sintetizadas a partir de imagens da base IMPA Faces(MENA-CHALCO; JUNIOR; VELHO, 2008).

5.7), nenhuma das imagens sintéticas (Figuras 37d, 37e e 37f) possui o enrugamento na bochecha.

A imagem apresentada na figura 38f apresenta uma leve deformação não natural do nariz e da boca para o lado esquerdo da imagem. Isso provavelmente aconteceu porque a imagem apresentada na figura 38c não está com posicionamento totalmente frontal. Desta maneira, a abordagem se beneficiaria de uma técnica de alinhamento capaz de reduzir esse problema.

Nenhuma das três imagens sintéticas da figura 38 resultou na região da sobrancelha levantada e uma grande abertura dos olhos, características comuns para a expressão da emoção de surpresa, como na imagem *ground truth*. Neste aspecto a abordagem proposta não foi capaz de alterar as posições dos *landmarks* dessa região de forma consistente. Uma possível explicação para esse fato é que o sistema não é capaz de identificar os *landmarks* da sobrancelha de maneira correta quando ela está muito levantada, o que é comum nessa expressão facial. Um exemplo do erro de detecção é apresentado na imagem da figura 39. Assim, uma maneira mais robusta de calcular o conjunto expressivo de *landmarks* faciais pode ajudar a contornar esse aspecto não expressivo.

As demais características expressivas (abertura da boca e levantamento do nariz) são encontradas nas imagens das figuras 38d e 38f, mas não na tabela 38e. A figura 38e pro-

Figura 38 – Exemplo de imagens sintéticas de surpresa para imagens selecionadas pelas abordagens características faciais, histograma de cor, e aleatória. À direita são apresentadas a face destino neutra e *ground truth*.

Fonte: Imagens sintetizadas a partir de imagens da base IMPA Faces(MENA-CHALCO; JUNIOR; VELHO, 2008).

Figura 39 – Exemplo de detecção de erro na posição dos *landmarks* faciais para região da sobrancelha.

Fonte: Imagens sintetizadas a partir de imagens da base IMPA Faces(MENA-CHALCO; JUNIOR; VELHO, 2008).

vavelmente não apresentou essas características expressivas devido à baixa expressividade da figura 38b. Além disso, um cálculo mais efetivo do conjunto expressivo de *landmarks* faciais poderia ajudar a amenizar esse problema.

A figura 40 apresenta exemplos de imagens sintetizadas pelas três diferentes abordagens de seleção das imagens fontes com variados indivíduos representando as seis expressões faciais das emoções consideradas universais. Para cada expressão facial são apresentas

imagens referentes a cada um das abordagens de seleção de características, além da imagem *ground truth*. Apesar das melhorias implementadas após o experimento da seção 6.1, alguns dos aspectos discutidos ainda são encontrados nas imagens.

As figuras 40a, 40i, 40q e 40r apresentam artefatos na região da testa. Tais artefatos são causados por diferenças de iluminação da região mapeadas durante o processo de síntese de enrugamentos (Seção 5.7). Nessas imagens é possível perceber que tais diferenças não estão representando aspectos como enrugamentos. Na verdade, esses artefatos são causados pela mudança de posição do cabelo. Artefatos desse tipo também são encontrados na região da bochecha na figura 40j e na região dos olhos na figura 40f. As melhorias implementadas (ver seção 6.1.5) apesar de reduzirem a quantidade de artefatos gerados, não conseguiram eliminar todos esses elementos não desejados. Uma solução para este problema poderia ser a detecção de *outliers* na imagem razão (AVERBUCH-ELOR et al., 2017).

Outro artefato comumente encontrado nas imagens sintéticas é a distorção de um ou mais componentes faciais. Nas figuras 40g, 40j e 40s há uma distorção assimétrica e não natural da região direita da face. Essas distorções provavelmente são provenientes de um cálculo do conjunto expressivo de *landmarks* (ver etapa dois da figura 25) de forma errônea. Uma possível explicação para esse erro é a mudança de perspectiva da face. Por exemplo, a pessoa pode ter movido o rosto levemente para a direita entre face neutra e a expressiva. Um alinhamento que considere esta mudança e um cálculo do conjunto expressivo de *landmarks* que consiga contornar esse problema poderiam gerar melhores resultados.

Também é possível perceber nas figuras 40b, 40k e 40w imagens com componentes faciais não naturais. Enquanto a figura 40b apresenta um olho extremamente pequeno, as figuras 40k e 40w apresentam bocas de tamanhos muito grandes. Esses componentes com tamanhos não naturais revelam uma deficiência do cálculo do conjunto expressivo de *landmarks* (ver etapa dois da figura 25) em preservar a identidade da face sintética.

Há também questões relacionadas à expressividade das emoções nas imagens sintetizadas. Por exemplo, as figuras 40g e 40w possuem pouquíssimos traços da expressão facial sintetizada. Esse aspecto pode ser resultado de dois principais fatores: i) as imagens fontes também possuem baixa expressividade e ii) o cálculo dos *landmarks* atenua, ainda mais, os traços expressivos. Outro exemplo relacionado à expressividade pode ser percebido nas imagens de aversão da figura 40. Nessas imagens as três figuras sintéticas (Figuras

Figura 40 – Exemplos de faces sintéticas para as três diferentes abordagens com variados indivíduos representando as seis expressões faciais consideradas universais.

	Características faciais	Histograma de cor	Aleatória	Ground truth
Aversão
	(a)	(b)	(c)	(d)
Medo
	(e)	(f)	(g)	(h)
Satisfação
	(i)	(j)	(k)	(l)
Raiva
	(m)	(n)	(o)	(p)
Surpresa
	(q)	(r)	(s)	(t)
Tristeza
	(u)	(v)	(w)	(x)

Fonte: Imagens retiradas ou sintetizadas a partir de imagens da base IMPA Faces(MENA-CHALCO; JUNIOR; VELHO, 2008).

40a, 40b e 40c) apresentam expressões faciais diferentes entre si e também diferentes da *ground truth* (Figura 40d). Essa diferença pode ser causada pela grande variedade de possíveis configurações que uma face de aversão possui. Afinal, uma mesma emoção pode ser expressa de diversas formas (EKMAN; FRIESEN, 1978).

Apesar de algumas imagens sintéticas apresentarem características não naturais e/ou não expressivas, as imagens sintetizadas a partir da similaridade por características faciais apresentaram, muitas vezes, resultados visualmente melhores. O que pode sugerir que faces similares à face destino podem melhorar os resultados. Ainda assim, como as imagens sintetizadas pelas três abordagens de seleção, incluindo as sintetizadas com faces similares por características faciais, apresentam fatores que diminuem o realismo, é difícil dizer se essas evidências visuais são suficientes para considerar que a abordagem é capaz de sintetizar imagens mais realistas. Entretanto, a melhoria das demais etapas da síntese de expressões faciais ajudaria a solucionar tais aspectos, que em sua maioria, ainda ficaram em aberto.

6.3 Experimento para a avaliação da busca por face similar

Nesta avaliação é verificado se as abordagens de seleção de imagens (Seções 4.2.2 e 5.5) são capazes de encontrar diferentes imagens faciais do mesmo indivíduo em diferentes condições. A ideia que embasa esta avaliação é que uma abordagem capaz de encontrar diferentes faces do mesmo indivíduo, deve conseguir encontrar faces similares. Idealmente, a pessoa mais parecida com ela seria ela própria.

Primeiramente, são calculadas a precisão e a revocação para cada imagem, como definido na seção 4.2.2. A figura 41 mostra o gráfico de precisão e revocação para quatro imagens da mesma pessoa. Por exemplo, buscando imagens similares à imagem *Heidi_Klum*₀₀₀₅ foram encontradas as imagens *Heidi_Klum*₀₀₀₅ e *Heidi_Klum*₀₀₀₄, o que garantiu precisão com valor um para os dois primeiros resultados. Depois foram encontradas as imagens *Princess_Caroline*₀₀₀₄, *Princess_Caroline*₀₀₀₃ e *Heidi_Klum*₀₀₀₂, assim a precisão para o terceiro resultado foi 0,6.

A precisão e revocação foi calculada para todas os indivíduos que possuíam pelo menos duas imagens, totalizando 1101 imagens de 175 indivíduos. Cada indivíduo possui um conjunto de duas a 145 fotos, sendo o mais comum um conjunto de três fotos de um mesmo indivíduo.

Figura 41 – Gráfico exemplificando a precisão e revocação para quatro imagens de um indivíduo com o descritor de histograma de cor.

Fonte: Rafael Luiz Testa, 2018

Para cada uma das curvas de precisão e revocação é calculada a área sob a curva. As médias da área sob a curva para as todas imagens em relação a cada um dos descritores foram 0,98 (DP=0,11) para características faciais, 0,20 (DP=0,16) para histograma de cores e 0,41 (DP=0,6) para aleatória. A comparação mostra uma área sobre a curva maior para o descritor das características faciais.

A figura 42 mostra o gráfico da precisão média em relação ao número de imagens. Por exemplo, caso um indivíduo possua um conjunto de cinco imagens no banco, os dados de precisão de cada uma de suas fotos foram considerados para calcular as médias de uma a cinco. Um outro indivíduo com um conjunto de seis imagens no banco contribui para o cálculo das médias de precisão de uma a seis imagens. Em geral, a precisão da métrica de características faciais se mostrou mais eficiente para encontrar as imagens, enquanto a similaridade por histograma de cor ficou bem próxima à aleatória.

6.4 Experimento para avaliação da síntese com face similar

Esta avaliação visa analisar, por meio de métricas objetivas, se a síntese de expressões faciais utilizando a face mais similar de acordo com as características faciais foi capaz de sintetizar imagens faciais com realismo adequado, quando comparada com outras abordagens de seleção das imagens fonte. Para isso, comparou-se o resultado esperado (*ground truth*) com o resultado obtido (imagem sintética) para as três diferentes maneiras

Figura 42 – Gráfico da precisão média por imagem encontrada para cada descritor.

Fonte: Rafael Luiz Testa, 2018

de encontrar uma face similar: comparando, as características faciais, histograma de cores ou aleatoriamente, conforme descrito na seção 4.2.2. As imagens foram avaliadas por três métricas *PSNR*, similaridade estrutural (Structural SIMilarity - *SSIM*) e diferenças entre *landmarks* (*ldif*).

As métricas *PSNR* e *SSIM* foram escolhidas por serem comumente utilizadas na comparação de imagens faciais (ver seção 3.5). Já a *ldif* foi escolhida por avaliar a estrutura facial, ao contrário das outras duas em que a avaliação faz uso dos *pixels* das imagens.

As *ldif* foi adaptada de Abboud, Davoine e Dang (2004). Ela compara algumas distâncias representativas em uma expressão facial, tais distâncias são ilustradas na imagem da figura 43. Para cada par de pontos faz-se a distância euclidiana entre eles. Então, são obtidos um vetor de características para cada imagem (sintética e *ground truth*). Cada posição do vetor armazena a distância entre as duas imagens. Os dois vetores são comparados também por meio da distância euclidiana. A diferença entre a métrica utilizada nesse trabalho e aquela apresentada em Abboud, Davoine e Dang (2004) é que neste trabalho é comparada a imagem referência e não a média dos *landmarks* por expressão facial.

A tabela 6 apresenta os resultados médios e os desvios padrão das três métricas mencionadas para os bancos de expressões faciais IMPA e MUG (ver seção 4.2.1). É possível perceber resultados similares para os três descritores em todas as métricas avaliadas para ambos bancos de expressões faciais. Se analisado sozinho, esse resultado poderia indicar

Figura 43 – Exemplo das distâncias utilizadas para compor a métrica de avaliação dos *landmarks* faciais.

Fonte: Adaptada de Abboud, Davoine e Dang (2004)

que a seleção das imagens fonte não exerce influência na expressão facial sintetizada. Entretanto, esse resultado destoa tanto da análise visual das imagens sintéticas, quanto dos resultados apresentados na seção anterior. Por exemplo, na análise visual (seção 6.2) a abordagem proposta para seleção de características, muitas vezes, foi capaz de preservar aspectos relacionados à preservação da identidade, pois não distorceu regiões da face, o que gera componentes faciais não naturais. Em muitos dos casos ela também foi capaz de sintetizar expressões com um realismo maior quando comparada com as outras abordagens (aleatória e histograma de cor).

Uma segunda hipótese é que o desempenho obtido esteja relacionado às métricas de avaliação utilizadas. A figura 44 apresenta uma imagem da diferença *pixel a pixel*, relacionada à métrica *PSNR*, e outra com o gradiente relacionado à métrica *SSIM* (AVANAKI, 2009). Nas imagens apresentadas na figura 43 são destacadas as mesmas regiões independentemente da imagem sintética para cada uma das abordagens de seleção, o que pode indicar que as métricas escolhidas não foram capazes de encontrar as diferenças desejadas. Apesar das métricas utilizadas serem provenientes da literatura, poucos trabalhos fizeram avaliação comparando com a imagem *ground truth*. Isso pode indicar que a comparação entre essas imagens faciais é ainda uma questão em aberto. A métrica baseada nos *landmarks* faciais proposta a partir de uma adaptação de Abboud, Davoine e Dang (2004) apresentou o maior desvio padrão, e poderia ser mais explorada com a proposição de métricas similares. Por exemplo, poderia haver uma adaptação de métricas comuns em outras áreas, na qual a avaliação da intersecção dos volumes de duas diferentes

Tabela 6 – Comparativo das áreas sob a curva da *ground truth* e imagem sintética para a seleção da imagens utilizando histograma de cor, características faciais e de forma aleatória nos bancos de expressão facial IMPA Faces e MUG.

IMPA			
	PSNR	SSIM	Landmarks
Aleatória	22,22 (DP=2,46)	0,767 (DP=0,052)	24,78 (DP=10,55)
Histograma de cor	21,98 (DP=2,45)	0,764 (DP=0,053)	24,82 (DP=11,51)
Características faciais	22,22 (DP=2,40)	0,762 (DP=0,051)	24,35 (DP=11,10)
MUG			
	PSNR	SSIM	Landmarks
Aleatória	21,72 (DP=2,18)	0,627 (DP=0,56)	27,57 (DP=16,41)
Histograma de cor	21,76 (DP=2,11)	0,628 (DP=0,055)	27,85 (DP=15,18)
Características faciais	21,82 (DP=2,15)	0,628 (DP=0,056)	29,10 (DP=17,06)

segmentações (UDUPA et al., 2006) poderia ser adaptada para comparar a forma dos componentes faciais.

6.5 Experimento para avaliação da preservação da identidade

Para avaliar a preservação da identidade foi realizado o mesmo teste apresentado na seção 6.4, entretanto comparou-se a imagem sintética com a face destino neutra. A ideia é avaliar a semelhança entre essas duas imagens e, assim, verificar se aspectos relacionados à identidade não foram perdidos.

A avaliação da seção 6.4 compara a imagem sintetizada com a imagem *ground truth*, e essa também é uma maneira de avaliar a preservação da identidade, já que a imagem *ground truth* pertence à mesma pessoa da F_{dn} . Entretanto, a avaliação apresentada nesta seção é importante, pois ao comparar a F_s com a F_{dn} , reduz-se a influência de outros aspectos, como a diferença de iluminação causada por outros fatores que não a expressão facial.

Na tabela 7 são apresentados os resultados para as métricas e os bancos de imagens supracitados. Como na seção 6.4, os valores encontrados tanto para média quanto para desvio padrão são bem próximos para as três abordagens de seleção das imagens fontes,

Figura 44 – Imagem diferença e imagem resultante do SSIM para um mesmo indivíduo considerando a diferença entre as imagen sintéticas (seleção das imagens fontes de modo aleatório, por histograma de cor e características faciais) e a imagem *ground truth*. A coluna *ground truth* mostra o resultado esperado para cada uma das imagens da linha correspondente.

Fonte: Imagens sintetizadas a partir de imagens da base MUG Faces(AIFANTI; PAPACHRISTOU; DELOPOULOS, 2010).

podendo indicar que, independente da imagem selecionada, as três abordagens foram capazes de preservar a identidade igualmente, ou que as métricas escolhidas não conseguiram encontrar diferenças significativas nesse aspecto, como discutido na seção 6.4.

De fato, esse resultado destoa do apresentado na seção 6.2, na qual componentes faciais como os olhos e o nariz têm o formato ou o tamanho diferente do natural, e também distorções em um dos lados da face. Esses aspectos levam a uma diferença maior da imagem original (F_{dn}), visto que os componentes faciais sintéticos destoam em tamanho e formato quando comparados à imagem original, e assim acabam por causar estranheza em relação à face sintetizada.

Tabela 7 – Comparativo das áreas sob a curva da imagem sintética e face destino neutra para a seleção da imagens utilizando histograma de cor, características faciais e de forma aleatória nos bancos de expressão facial IMPA Faces e MUG.

IMPA			
	PSNR	SSIM	Landmarks
Aleatória	26,34 (DP=2,53)	0,856 (DP=0,038)	23,44 (DP=11,74)
Histograma de cor	25,92 (DP=2,28)	0,850 (DP=0,034)	25,05 (DP=12,37)
Características faciais	25,39 (DP=2,24)	0,843 (DP=0,038)	24,05 (DP=12,07)
MUG			
	PSNR	SSIM	Landmarks
Aleatória	25,97 (DP=2,41)	0,738 (DP=0,055)	36,87 (DP=20,72)
Histograma de cor	26,08 (DP=2,33)	0,739 (DP=0,051)	37,65 (DP=20,79)
Características faciais	25,89 (DP=2,36)	0,737 (DP=0,051)	38,35 (DP=21,34)

Um exemplo de uma face que perdeu as características relativas à identidade é ilustrado na imagem da figura 40w, em que a região da boca é grande quando comparada à imagem da figura 40x, e o mesmo se percebe ao comparar a figura 40k com a 40i. As face com distorções no rosto são mais evidentes nas imagens da figuras 40g e 40s. A análise visual dessas imagens indica que a seleção aleatória das imagens fontes leva a mais distorções que as outras duas abordagens. Assim, ressalta-se a importância de selecionar uma imagem fonte similar para aumentar a percepção da preservação da identidade.

7 Conclusões e trabalhos futuros

O presente trabalho propôs uma nova abordagem para a síntese realista das expressões faciais que considera faces similares. A primeira contribuição resultante desse trabalho foi estabelecer uma visão completa da área por meio da RS da literatura, tanto para a síntese baseada em “Deformação e Mapeamento” quanto para as baseadas em “Aprendizado de Máquina”. Apesar de ter sido encontrado um grande número de abordagens baseadas em aprendizado de máquina, muitas das abordagens são mistas. Provavelmente a quantidade reduzida de imagens nos bancos de expressões faciais dificultam a proposição de abordagens puras baseadas em “Aprendizado de Máquina” para a síntese.

Esse trabalho também gerou contribuições na proposição de uma nova abordagem para síntese das expressões faciais, mais especificamente em relação às abordagens de “Deformação e Mapeamento”. A maior contribuição resultante deste trabalho foi a proposição do uso de *ResNet* para extrair características faciais mais relacionadas à identidade para encontrar pessoas similares. A análise visual mostra que as faces similares selecionadas contribuíram para uma melhoria de aspectos relacionados ao realismo, como componentes faciais com tamanhos naturais e a redução dos artefatos gerados.

O trabalho ainda obteve outras contribuições em outros aspectos da síntese de expressões faciais por “Deformação e Mapeamento”. A triangulação manual dos *landmarks* faciais diminui a chance de artefatos na região interna da boca. O mascaramento das regiões que possuem enrugamentos ajuda a sintetizar imagens com menos artefatos independente das mudanças de iluminação geradas por aspectos não relacionados às expressões faciais.

O método proposto juntamente com as contribuições destacadas foram capazes de sintetizar imagens com um realismo adequado para várias das possíveis aplicações. Apesar do realismo obtido, as imagens sintéticas finais ainda apresentam aspectos que diminuem a percepção de realismo. Assim, como trabalho futuro sugere-se a melhoria na identificação dos *landmarks* faciais (especialmente para configurações faciais extremas como surpresa), no cálculo do conjunto expressivo de *landmarks*, na adaptação das condições de iluminação para componentes de outras faces (região interna da boca) e na proposição de técnicas que transferiam apenas as diferenças de iluminação relacionadas à expressão facial.

Além disso, o trabalho também propôs três maneiras experimentais de avaliar as imagens sintéticas, sendo duas objetivas e uma subjetiva. As objetivas visam avaliar o quão

as imagens sintéticas se distanciaram de um exemplo esperado. As métricas utilizadas para comparação das imagens, comumente utilizadas na literatura, não se mostram robustas o suficiente para avaliar as imagens. Portanto, a proposição de técnicas que consigam comparar imagens faciais expressivas ainda é uma questão em aberto.

A avaliação subjetiva se mostrou eficiente em avaliar as imagens sintéticas. Além disso, nessa avaliação também foram levantados os aspectos que contribuem para o aumento e diminuição da percepção de realismo em imagens sintéticas. Enquanto fatores como bordas com irregularidades diferenças de cor, e componentes faciais de tamanhos não naturais contribuem para uma menor percepção de realismo; outros fatores como enrugamentos, preservação da identidade e expressividade aumentam a percepção de realismo nas imagens. Destaca-se, ainda, a necessidade de experimentos com usuários para avaliação específica desses aspectos como a preservação da identidade e a expressividade.

Os resultados obtidos constituem uma contribuição para a área de processamento de imagens ao propor novas técnicas para síntese e avaliação de expressões faciais, e também para as diversas áreas de aplicação nas quais os resultados poderão ser aplicados.

Por isso, além do impacto acadêmico evidenciado, as imagens sintéticas resultantes desse trabalho se mostram viáveis, em muitos casos, para serem utilizadas no contexto de diagnóstico e treinamento do reconhecimento das emoções faciais para pessoas com distúrbios relacionados a essas habilidades. A abordagem é capaz de sintetizar imagens de diferentes níveis dependendo da imagem fonte utilizada, possibilitando uma maior personalização do tratamento para que melhor se adéque às necessidades de cada paciente.

Também é esperado um impacto econômico com a redução do tempo requerido do profissional de saúde para o preparo de imagens personalizadas para um determinado tratamento. Além disso, essas imagens poderão ser utilizadas em outras áreas, como animação e entretenimento. Portanto, esse projeto também pode resultar em uma contribuição econômica nessas áreas ao permitir a automação da geração das emoções faciais em avatares.

Por fim, não foram encontradas ferramentas gratuitas para síntese das expressões faciais, portanto, a disponibilização desse *software* resulta em uma contribuição tecnológica.

Referências¹

ABBOUD, B.; DAVOINE, F.; DANG, M. Facial expression recognition and synthesis based on an appearance model. Signal Processing: Image Communication, v. 19, n. 8, p. 723 – 740, 2004. ISSN 0923-5965. Citado 2 vezes nas páginas 90 e 91.

AIFANTI, N.; PAPACHRISTOU, C.; DELOPOULOS, A. The mug facial expression database. In: 11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10. [S.l.: s.n.], 2010. p. 1–4. ISSN 2158-5873. Citado 2 vezes nas páginas 51 e 93.

AMBERG, B.; KNOTHE, R.; VETTER, T. Expression invariant 3d face recognition with a morphable model. In: 2008 8th IEEE International Conference on Automatic Face Gesture Recognition. [S.l.: s.n.], 2008. p. 1–6. Citado na página 18.

ANDREWS, J. H.; BRIAND, L. C.; LABICHE, Y. Is mutation an appropriate tool for testing experiments? In: Proceedings of the 27th International Conference on Software Engineering. New York, NY, USA: ACM, 2005. (ICSE '05), p. 402–411. ISBN 1-58113-963-2. Citado na página 45.

AVANAKI, A. N. Exact global histogram specification optimized for structural similarity. Optical Review, v. 16, n. 6, p. 613–621, Nov 2009. ISSN 1349-9432. Disponível em: <https://doi.org/10.1007/s10043-009-0119-z>. Citado na página 91.

AVERBUCH-ELOR, H. et al. Bringing portraits to life. ACM Trans. Graph., ACM, New York, NY, USA, v. 36, n. 6, p. 196:1–196:13, nov. 2017. ISSN 0730-0301. Disponível em: <http://doi.acm.org/10.1145/3130800.3130818>. Citado 11 vezes nas páginas 10, 58, 66, 67, 69, 70, 72, 73, 74, 78 e 86.

AZEVEDO, E.; CONCI, A. Computação gráfica: geração de imagens. [S.l.]: Elsevier, 2003. v. 1. ISBN 9788535212525. Citado 2 vezes nas páginas 22 e 23.

BAGGIO, D. L. et al. Mastering OpenCV 3 - Second Edition. 2nd. ed. [S.l.]: Packt Publishing, 2017. ISBN 1786467178, 9781786467171. Citado na página 61.

BENGIO, Y. Learning deep architectures for ai. Found. Trends Mach. Learn., Now Publishers Inc., Hanover, MA, USA, v. 2, n. 1, p. 1–127, jan. 2009. ISSN 1935-8237. Disponível em: <http://dx.doi.org/10.1561/22000000006>. Citado na página 30.

BERG, M. d. et al. Computational Geometry: Algorithms and Applications. 3rd ed.. ed. Santa Clara, CA, USA: Springer-Verlag TELOS, 2008. ISBN 3540779736, 9783540779735. Citado na página 28.

CHENG, Y.; LING, S. 3d animated facial expression and autism in taiwan. In: IEEE. Advanced Learning Technologies, 2008. ICALT'08. Eighth IEEE International Conference on. [S.l.], 2008. p. 17–19. Citado na página 17.

CLARK, T. F.; WINKIELMAN, P.; MCINTOSH, D. N. Autism and the extraction of emotion from briefly presented facial expressions: stumbling at the first step of empathy. Emotion, American Psychological Association, v. 8, n. 6, p. 803, 2008. Citado na página 17.

¹ De acordo com a Associação Brasileira de Normas Técnicas. NBR 6023.

- COCHRAN, W. G. The χ^2 test of goodness of fit. Ann. Math. Statist., The Institute of Mathematical Statistics, v. 23, n. 3, p. 315–345, 09 1952. Disponível em: <https://doi.org/10.1214/aoms/1177729380>. Citado na página 55.
- CONCI, A.; AZEVEDO, E.; LETA, F. Computação gráfica: teoria e prática. [S.l.]: Elsevier, 2008. v. 2. ISBN 9788535223293. Citado na página 24.
- CRISCUOLO, M. SlimRank: um modelo de seleção de respostas para perguntas de consumidores. Tese (Doutorado) — Universidade de São Paulo, 2017. Citado na página 31.
- CUN, Y. L. et al. Advances in neural information processing systems 2. In: TOURETZKY, D. S. (Ed.). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1990. cap. Handwritten Digit Recognition with a Back-propagation Network, p. 396–404. ISBN 1-55860-100-7. Disponível em: <http://dl.acm.org/citation.cfm?id=109230.109279>. Citado na página 30.
- DENG, J. et al. ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09. [S.l.: s.n.], 2009. Citado na página 58.
- DYCK, M. et al. Recognition profile of emotions in natural and virtual faces. PLOS ONE, Public Library of Science, v. 3, n. 11, p. 1–8, 11 2008. Disponível em: <https://doi.org/10.1371/journal.pone.0003628>. Citado na página 77.
- EKMAN, P.; FRIESEN, W. V. Constants across cultures in the face and emotion. Journal of personality and social psychology, American Psychological Association, v. 17, n. 2, p. 124, 1971. Citado na página 16.
- EKMAN, P.; FRIESEN, W. V. Facial Action Coding System: A Technique for the Measurement of Facial Movement. Palo Alto: Consulting Psychologists Press, 1978. Citado na página 88.
- EKMAN, P.; FRIESEN, W. V.; HAGER, J. C. Facs investigator's guide. A human face, 2002. Citado 2 vezes nas páginas 70 e 82.
- EKMAN, P.; HAGER, J. C.; FRIESEN, W. V. The symmetry of emotional and deliberate facial actions. Psychophysiology, Blackwell Publishing Ltd, v. 18, n. 2, p. 101–106, 1981. ISSN 1469-8986. Disponível em: <http://dx.doi.org/10.1111/j.1469-8986.1981.tb02919.x>. Citado na página 17.
- EKMAN, P.; ROSENBERG, E. L. What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS). [S.l.]: Oxford University Press, USA, 1997. Citado na página 44.
- EVERINGHAM, M. et al. The pascal visual object classes challenge: A retrospective. Int. J. Comput. Vision, Kluwer Academic Publishers, Hingham, MA, USA, v. 111, n. 1, p. 98–136, jan. 2015. ISSN 0920-5691. Disponível em: <http://dx.doi.org/10.1007/s11263-014-0733-5>. Citado na página 58.
- FASEL, B.; LUETTIN, J. Automatic facial expression analysis: a survey. Pattern Recognition, v. 36, n. 1, p. 259 – 275, 2003. ISSN 0031-3203. Citado 2 vezes nas páginas 22 e 23.

- FOLEY, J. D. Computer graphics: Principles and practice, in C. [S.l.]: Addison-Wesley Professional, 1996. v. 12110. Citado na página 22.
- FOLEY, J. D. et al. Computer Graphics: Principles and Practice (3rd Ed.). Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1990. ISBN 0-201-12110-7. Citado na página 23.
- FOLK, M. et al. An overview of the hdf5 technology suite and its applications. In: Proceedings of the EDBT/ICDT 2011 Workshop on Array Databases. New York, NY, USA: ACM, 2011. (AD '11), p. 36–47. ISBN 978-1-4503-0614-0. Disponível em: <http://doi.acm.org/10.1145/1966895.1966900>. Citado na página 65.
- FUJISHIRO, H. et al. A natural smile synthesis from an artificial smile. In: SIGGRAPH '09: Posters. New York, NY, USA: ACM, 2009. (SIGGRAPH '09), p. 59:1–59:1. Disponível em: <http://doi.acm.org/10.1145/1599301.1599360>. Citado na página 18.
- GLOROT, X.; BENGIO, Y. Understanding the difficulty of training deep feedforward neural networks. In: TEH, Y. W.; TITTERINGTON, M. (Ed.). Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Chia Laguna Resort, Sardinia, Italy: PMLR, 2010. (Proceedings of Machine Learning Research, v. 9), p. 249–256. Disponível em: <http://proceedings.mlr.press/v9/glorot10a.html>. Citado na página 32.
- GOLAN, O.; BARON-COHEN, S. Systemizing empathy: Teaching adults with asperger syndrome or high-functioning autism to recognize complex emotions using interactive multimedia. Development and Psychopathology, null, p. 591–617, 6 2006. ISSN 1469-2198. Citado na página 17.
- GONÇALVES, V. M.; DELAMARO, M. E.; NUNES, F. L. S. Applying graphical oracles to evaluate image segmentation results. Journal of the Brazilian Computer Society, v. 23, n. 1, p. 1, Jan 2017. ISSN 1678-4804. Citado 2 vezes nas páginas 27 e 28.
- GOSSELIN, P.; KIROUAC, G. et al. Decoding facial emotional prototypes. Canadian journal of experimental psychology, v. 49, n. 3, p. 313, 1995. Citado na página 16.
- GRYNSZPAN, O.; MARTIN, J.-C.; NADEL, J. Multimedia interfaces for users with high functioning autism: An empirical investigation. International Journal of Human-Computer Studies, Elsevier, v. 66, n. 8, p. 628–639, 2008. Citado na página 17.
- HARMS, M. B.; MARTIN, A.; WALLACE, G. L. Facial emotion recognition in autism spectrum disorders: a review of behavioral and neuroimaging studies. Neuropsychology review, Springer, v. 20, n. 3, p. 290–322, 2010. Citado na página 16.
- HE, K. et al. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.: s.n.], 2016. p. 770–778. ISSN 1063-6919. Citado 2 vezes nas páginas 32 e 64.
- HEARN, D. D.; BAKER, M. P. Computer Graphics with OpenGL. 3. ed. [S.l.]: Prentice Hall Professional Technical Reference, 2003. ISBN 0130153907. Citado na página 29.
- HESS, U. The communication of emotion. Emotions, qualia and consciousness, p. 397–409, 2001. Citado na página 16.

- IMPETT, L.; ROBINSON, P.; BALTRUSAITIS, T. A facial affect mapping engine. In: Proceedings of the Companion Publication of the 19th International Conference on Intelligent User Interfaces. New York, NY, USA: ACM, 2014. (IUI Companion '14), p. 33–36. ISBN 978-1-4503-2729-9. Citado na página 70.
- KANADE, T.; COHN, J. F.; TIAN, Y. Comprehensive database for facial expression analysis. In: Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on. [S.l.: s.n.], 2000. p. 46–53. Citado 2 vezes nas páginas 42 e 44.
- KAZEMI, V.; SULLIVAN, J. One millisecond face alignment with an ensemble of regression trees. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Washington, DC, USA: IEEE Computer Society, 2014. (CVPR '14), p. 1867–1874. ISBN 978-1-4799-5118-5. Citado 3 vezes nas páginas 27, 58 e 65.
- KEEVE, E. et al. Deformable modeling of facial tissue for craniofacial surgery simulation. Computer Aided Surgery, Taylor & Francis, v. 3, n. 5, p. 228–238, 1998. Citado na página 18.
- KING, D. E. Max-margin object detection. CoRR, abs/1502.00046, 2015. Disponível em: <http://arxiv.org/abs/1502.00046>. Citado 2 vezes nas páginas 25 e 26.
- KING, D. E. DLIB face detection dataset. 2016. http://dlib.net/files/data/dlib_face_detection_dataset-2016-09-30.tar.gz. [Online; Acessado em: 01 de outubro de 2018]. Citado na página 58.
- KING, D. E. High Quality Face Recognition with Deep Metric Learning. 2017. <http://blog.dlib.net/2017/02/high-quality-face-recognition-with-deep.html>. [Online; Acessado em: 01 de outubro de 2018]. Citado na página 65.
- KOESTINGER PAUL WOHLHART, P. M. R. M.; BISCHOF, H. Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization. In: Proc. First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies. [S.l.: s.n.], 2011. Citado na página 58.
- LAHIRI, U. et al. Design of a virtual reality based adaptive response technology for children with autism. Neural Systems and Rehabilitation Engineering, IEEE Transactions on, IEEE, v. 21, n. 1, p. 55–64, 2013. Citado na página 17.
- LECUN, Y.; BENGIO, Y.; HINTON, G. Deep learning. Nature, Nature Publishing Group, a division of Macmillan Publishers Limited. All Rights Reserved. SN -, v. 521, p. 436 EP –, May 2015. Disponível em: <https://doi.org/10.1038/nature14539>. Citado 3 vezes nas páginas 30, 31 e 32.
- LI, K. et al. A data-driven approach for facial expression retargeting in video. IEEE Transactions on Multimedia, v. 16, n. 2, p. 299–310, Feb 2014. ISSN 1520-9210. Citado na página 64.
- LI, X.; CHANG, C.-C.; CHANG, S.-K. Face alive icon. Journal of Visual Languages & Computing, Elsevier, v. 18, n. 4, p. 440–453, 2007. ISSN 1045-926X. Visual Interactions in Software Artifacts. Citado na página 18.

LI, Z.; ZHU, C.; GOLD, C. Digital terrain modeling: principles and methodology. [S.l.]: CRC press, 2004. Citado na página 61.

LIU, C.; YUEN, J.; TORRALBA, A. Sift flow: Dense correspondence across scenes and its applications. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 33, n. 5, p. 978–994, May 2011. ISSN 0162-8828. Citado na página 44.

LIU, Z.; SHAN, Y.; ZHANG, Z. Expressive expression mapping with ratio images. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. New York, NY, USA: ACM, 2001. (SIGGRAPH '01), p. 271–276. ISBN 1-58113-374-X. Citado 2 vezes nas páginas 68 e 76.

LUCEY, P. et al. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. [S.l.: s.n.], 2010. p. 94–101. ISSN 2160-7508. Citado na página 42.

LYONS, M. et al. Coding facial expressions with gabor wavelets. In: Proceedings of the 3rd. International Conference on Face & Gesture Recognition. Washington, DC, USA: IEEE Computer Society, 1998. (FG '98), p. 200–. ISBN 0-8186-8344-9. Citado na página 42.

MAGNOR, M. A. Video-Based Rendering. [S.l.]: AK Peters Ltd, 2005. ISBN 1568812442. Citado na página 23.

MARČETIĆ, D.; SOLDIĆ, M.; RIBARIĆ, S. Hybrid cascade model for face detection in the wild based on normalized pixel difference and a deep convolutional neural network. In: FELSBERG, M.; HEYDEN, A.; KRÜGER, N. (Ed.). Computer Analysis of Images and Patterns. Cham: Springer International Publishing, 2017. p. 379–390. ISBN 978-3-319-64698-5. Citado na página 58.

MARSCHNER, S. R.; GREENBERG, D. P. Inverse lighting for photography. In: SOCIETY FOR IMAGING SCIENCE AND TECHNOLOGY. Color and Imaging Conference. [S.l.], 1997. v. 1997, n. 1, p. 262–265. Citado na página 30.

MENA-CHALCO, J.; JUNIOR, R. C.; VELHO, L. Banco de dados de faces 3D: IMPA-FACE3D. [S.l.], 2008. 8 p. Disponível em: <http://app.visgraf.impa.br/database/faces/>. Citado 15 vezes nas páginas 51, 54, 67, 69, 75, 76, 77, 78, 79, 80, 82, 83, 84, 85 e 87.

MENDI, E.; BAYRAK, C. Facial animation framework for web and mobile platforms. In: 2011 IEEE 13th International Conference on e-Health Networking, Applications and Services. [S.l.: s.n.], 2011. p. 52–55. Citado na página 18.

MOHAMMADIAN, A.; AGHAEINIA, H.; TOWHIDKHAH, F. Diverse videos synthesis using manifold-based parametric motion model for facial understanding. IET Image Processing, v. 10, n. 4, p. 253–260, 2016. ISSN 1751-9659. Citado 3 vezes nas páginas 18, 39 e 48.

MOURÃO-MIRANDA, J. et al. Pattern recognition and functional neuroimaging help to discriminate healthy adolescents at risk for mood disorders from low risk adolescents. PloS one, Public Library of Science, v. 7, n. 2, p. e29482, 2012. Citado na página 16.

NEWSCHAFER, C. J.; CURRAN, L. K. Autism: an emerging public health problem. Public health reports, Association of Schools of Public Health, v. 118, n. 5, p. 393, 2003. Citado na página 17.

NG, H.; WINKLER, S. A data-driven approach to cleaning large face datasets. In: 2014 IEEE International Conference on Image Processing (ICIP). [S.l.: s.n.], 2014. p. 343–347. ISSN 1522-4880. Citado 3 vezes nas páginas 51, 58 e 65.

NOH, J.-y.; NEUMANN, U. Expression cloning. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. New York, NY, USA: ACM, 2001. (SIGGRAPH '01), p. 277–288. ISBN 1-58113-374-X. Citado na página 18.

PARKHI, O. M.; VEDALDI, A.; ZISSERMAN, A. Deep face recognition. In: British Machine Vision Conference. [S.l.: s.n.], 2015. Citado 3 vezes nas páginas 51, 58 e 65.

PATTERSON, J.; GIBSON, A. Deep Learning: A Practitioner's Approach. [S.l.]: "O'Reilly Media, Inc.", 2017. Citado na página 31.

PEDRINI, H.; SCHWARTZ, W. Análise de imagens digitais: princípios, algoritmos e aplicações. [S.l.]: THOMSON PIONEIRA, 2008. ISBN 9788522105953. Citado 4 vezes nas páginas 22, 23, 29 e 33.

PETITCREW, M.; ROBERTS, H. Systematic reviews in the social sciences: A practical guide. [S.l.]: John Wiley & Sons, 2008. Citado na página 35.

PIERCE, K. et al. The brain response to personally familiar faces in autism: findings of fusiform activity and beyond. Brain, Oxford University Press, v. 127, n. 12, p. 2703–2716, 2004. Citado na página 17.

RADEMACHER, P. M. Measuring the Perceived Visual Realism of Images. Tese (Doutorado) — University of North Carolina, 2002. Citado na página 55.

REINHARD, E. et al. High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting (The Morgan Kaufmann Series in Computer Graphics). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005. ISBN 0125852630. Citado na página 30.

ROCCA, C. C. d. A. et al. Facial emotion recognition in bipolar disorder: a critical review. Revista Brasileira de Psiquiatria, SciELO Brasil, v. 31, n. 2, p. 171–180, 2009. Citado na página 16.

SAGONAS, C. et al. 300 faces in-the-wild challenge: database and results. Image and Vision Computing, v. 47, p. 3 – 18, 2016. ISSN 0262-8856. 300-W, the First Automatic Facial Landmark Detection in-the-Wild Challenge. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0262885616000147>. Citado na página 58.

SCHERER, D.; MÜLLER, A.; BEHNKE, S. Evaluation of pooling operations in convolutional architectures for object recognition. In: DIAMANTARAS, K.; DUCH, W.; ILIADIS, L. S. (Ed.). Artificial Neural Networks – ICANN 2010. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p. 92–101. ISBN 978-3-642-15825-4. Citado na página 31.

- SCHERER, K. R. Vocal communication of emotion: A review of research paradigms. Speech communication, Elsevier, v. 40, n. 1, p. 227–256, 2003. Citado na página 16.
- SEYAMA, J.; NAGAYAMA, R. S. The uncanny valley: Effect of realism on the impression of artificial human faces. Presence: Teleoperators and Virtual Environments, v. 16, n. 4, p. 337–351, 2007. Disponível em: <https://doi.org/10.1162/pres.16.4.337>. Citado na página 75.
- SHUM, H.-Y.; CHAN, S.-C.; KANG, S. B. Image-based rendering. [S.l.]: Springer Science & Business Media, 2008. Citado 2 vezes nas páginas 24 e 25.
- SIDDIQUI, J. R. et al. Towards visual mapping in industrial environments - a heterogeneous task-specific and saliency driven approach. In: 2016 IEEE International Conference on Robotics and Automation (ICRA). [S.l.: s.n.], 2016. p. 5766–5773. Citado na página 25.
- SJOEBERG, D. I. K. et al. A survey of controlled experiments in software engineering. IEEE Transactions on Software Engineering, v. 31, n. 9, p. 733–753, Sept 2005. ISSN 0098-5589. Citado na página 45.
- SONG, M. et al. Image ratio features for facial expression recognition application. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), v. 40, n. 3, p. 779–788, June 2010. ISSN 1083-4419. Citado na página 18.
- TAYLOR, S. F.; MACDONALD, A. W. Brain mapping biomarkers of socio-emotional processing in schizophrenia. Schizophrenia bulletin, MPRC, v. 38, n. 1, p. 73–80, 2012. Citado na página 16.
- TESTA, R. L. et al. Synthesis of facial expressions in photographs: characteristics, approaches, and challenges. ACM Comput. Surv., ACM, New York, NY, USA, 2018. ISSN 0360-0300. (in press). Citado 3 vezes nas páginas 35, 37 e 38.
- TESTA, R. L. et al. Factors influencing the perception of realism in synthetic facial expression. In: 2018 31th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI). [S.l.: s.n.], 2018. p. 127–133. (in press). Citado 6 vezes nas páginas 54, 72, 75, 76, 77 e 78.
- THELIN, T.; RUNESON, P.; WOHLIN, C. An experimental comparison of usage-based and checklist-based reading. IEEE Transactions on Software Engineering, v. 29, n. 8, p. 687–704, Aug 2003. ISSN 0098-5589. Citado na página 45.
- UDUPA, J. K. et al. A framework for evaluating image segmentation algorithms. Computerized Medical Imaging and Graphics, v. 30, n. 2, p. 75 – 87, 2006. ISSN 0895-6111. Disponível em: <http://www.sciencedirect.com/science/article/pii/S089561110500114X>. Citado na página 92.
- WANG, N. et al. Facial feature point detection: A comprehensive survey. Neurocomputing, p. –, 2017. ISSN 0925-2312. Citado 4 vezes nas páginas 26, 27, 49 e 58.
- WANG, Z.; BOVIK, A. C. Modern image quality assessment. Synthesis Lectures on Image, Video, and Multimedia Processing, Morgan & Claypool Publishers, v. 2, n. 1, p. 1–156, 2006. Citado na página 33.

WINKLER, S. Digital video quality: vision models and metrics. [S.l.]: John Wiley & Sons, 2005. Citado na página 33.

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. Citado na página 45.

XIE, W. et al. Lighting difference based wrinkle mapping for expression synthesis. In: 2015 8th International Congress on Image and Signal Processing (CISP). [S.l.: s.n.], 2015. p. 636–641. Citado na página 64.

XIONG, L. et al. Eye synthesis using the eye curve model. Image and Vision Computing, v. 28, n. 3, p. 329 – 342, 2010. ISSN 0262-8856. Citado 5 vezes nas páginas 19, 39, 44, 45 e 48.

XIONG, L. et al. Eye synthesis using the eye curve model. In: 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007). [S.l.: s.n.], 2007. v. 2, p. 531–534. ISSN 1082-3409. Citado 2 vezes nas páginas 39 e 48.

YANG, S. et al. Wider face: A face detection benchmark. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.: s.n.], 2016. Citado na página 58.

ZAFEIRIOU, S.; ZHANG, C.; ZHANG, Z. A survey on face detection in the wild: Past, present and future. Computer Vision and Image Understanding, v. 138, p. 1 – 24, 2015. ISSN 1077-3142. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1077314215000727>. Citado na página 49.

ZHANG, Y. et al. Facial expression cloning with elastic and muscle models. Journal of Visual Communication and Image Representation, v. 25, n. 5, p. 916 – 927, 2014. ISSN 1047-3203. Citado 2 vezes nas páginas 18 e 45.

ZHANG, Z.; SONG, Y.; QI, H. Age progression/regression by conditional adversarial autoencoder. CoRR, abs/1702.08423, 2017. Disponível em: <http://arxiv.org/abs/1702.08423>. Citado na página 18.

ZHOU, Y.; SHI, B. E. Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII). [S.l.: s.n.], 2017. p. 370–376. Citado 2 vezes nas páginas 19 e 44.