Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Métodos para cálculo de razão de verossimilhança para utilização de sistemas de reconhecimento facial em cenários forenses

Rafael Oliveira Ribeiro

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Orientador
Prof. Dr. Flávio de Barros Vidal

Brasília
2023

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

Oliveira Ribeiro, Rafael
OR484m Métodos para cálculo de razão de verossimilhança para utilização de sistemas de reconhecimento facial em cenários forenses / Rafael Oliveira Ribeiro; orientador Flávio de Barros Vidal. -- Brasília, 2023.
89 p.

Dissertação(Mestrado em Informática) -- Universidade de Brasília, 2023.

1. Reconhecimento Facial. 2. Ciências Forenses. 3. Interpretação de Evidência. 4. Razão de Verossimilhança. I. de Barros Vidal, Flávio, orient. II. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Métodos para cálculo de razão de verossimilhança para utilização de sistemas de reconhecimento facial em cenários forenses

Rafael Oliveira Ribeiro

Dissertação apresentada como requisito parcial para
conclusão do Mestrado em Informática

Prof. Dr. Flávio de Barros Vidal (Orientador)
Departamento de Ciência da Computação - UnB

Prof. Dr. David Menotti Gomes
Departamento de Informática - UFPR

Prof. Dr. João Carlos Raposo Neves
Departamento de Informática - UBI, Portugal

Prof. Dr. Díbio Leandro Borges (Suplente)
Departamento de Ciência da Computação - UnB

Prof. Dr. Ricardo Pezzuol Jacobi
Coordenador do Programa de Pós-graduação em Informática

Brasília, 19 de junho de 2023

Dedicatória

Dedico este trabalho a meus pais, Luiz Alberto e Brandaly, que sempre me estimularam a estudar e a perseguir meus sonhos. Dedico também à minha esposa e filhos, Carolinne, Felipe e Henrique, pelo apoio e pela compreensão pelas ausências durante a pesquisa.

Agradecimentos

Agradeço ao meu orientador, Prof. Dr. Flávio de Barros Vidal, pelo suporte e pela compreensão nos momentos de difícil conciliação entre o trabalho e a pesquisa. Agradeço ao Dr. Arnout Ruifrok e ao Dr. João Neves pela colaboração e apoio nos experimentos relacionados a agregação de *embeddings*. Agradeço também aos colegas Gustavo, Janine e Paulo Max pelas discussões enriquecedoras no grupo AutoFaceRec. Agradeço por fim aos meus chefes imediatos, extensivo às instâncias superiores, durante o período em que estive dedicado a esta pesquisa, pela compreensão e apoio para conciliação de horários.

Resumo

Na área forense, o exame pericial de comparação facial tem adquirido maior relevância à medida em que cresce o número de dispositivos com capacidade de gravação de imagens e, por conseguinte, aumenta o número de crimes em que os autores têm suas faces capturadas em imagens. Atualmente esse exame pericial é baseado na análise e comparação manual de elementos morfológicos da face e os resultados são expressos de forma qualitativa, o que dificulta a sua reprodutibilidade e a combinação de seus resultados com outras evidências pela instância julgadora. Este trabalho tem como objetivo avaliar métodos para expressar os resultados do exame de forma quantitativa, com o cálculo de razão de verossimilhança (do inglês *Likelihood-Ratio* – LR) a partir de escores obtidos de sistemas de reconhecimento facial. Além de facilitar a reprodutibilidade dos resultados, aspecto crítico na área forense, os métodos avaliados permitem a validação empírica de desempenho nas condições de cada caso. Neste trabalho foram avaliados métodos paramétricos e não-paramétricos para cálculo de LR a partir de escores, utilizando dois sistemas de reconhecimento facial de código aberto, ArcFace e FaceNet, e cinco bases com imagens faciais representativas de cenários frequentemente encontrados em casos periciais: imagens de mídias sociais e de câmeras de CFTV. Além disso, foram realizados experimentos relacionados à agregação de *embeddings* em casos onde há mais de uma imagem do indivíduo de interesse. Estes experimentos demonstraram melhora substancial no cálculo de LR a partir de sistemas de reconhecimento facial, especialmente nos cenários envolvendo imagens de pior qualidade: redução na C_{llr} em até 95% (de 0,249 para 0,012) para imagens de CFTV e de até 96% (de 0,083 para 0,003) para imagens de mídias sociais.

Palavras-chave: reconhecimento facial, ciências forenses, interpretação de evidência, razão de verossimilhança

Abstract

Forensic face comparison is becoming more relevant as the number of devices with image recording capabilities increase, with a consequential increase in the number of crimes in which the face of the perpetrator is recorded. This forensic examination is still based on the manual analysis and comparison of morphological features of the faces. Its results are expressed qualitatively, making it difficult to reproduce and combine with other evidence. This work evaluates methods to obtain a quantitative result for the examination, with the computation of score-based Likelihood-Ratio - LR. Face recognition systems are used to obtain scores that are then converted to an LR. The methods investigated in this work facilitate reproducibility, a critical aspect in forensics, and it also allow for the empirical validation of performance in the conditions of each forensic case. We evaluate parametric and non-parametric methods for LR computation. Two open-source face recognition models were used (ArcFace and FaceNet) on images from five datasets that are representative of common scenarions in forensic casework: images from social media and images from CCTV cameras. We also investigate strategies for embedding aggregation in cases where there is more than one image of the person of interest. These experiments demonstrate substantial improvements in forensic evaluation settings, with improvements in C_{llr} of up to 95% (from 0.249 to 0.012) for CCTV images and of up to 96% (from 0.083 to 0.003) for social media images.

Keywords: face recognition, forensic science, evaluation of evidence, likelihood ratio

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Símbolos

C_{llr} *Log-Likelihood-Ratio Cost.*

AUC Área sob a curva ROC.

BFW *Balanced Faces in the Wild.*

BSV *Between-Sources Variability.*

CASIA *Institute of Automation of the Chinese Academy of Sciences.*

CFTV Circuito Fechado de Televisão.

CS *Confusion Score.*

DNA ácido desoxirribonucleico.

ECE *Empirical Cross-Entropy.*

ENFSI *European Network of Forensic Science Institutes.*

EU *Expected Utility.*

FEI Faculdade de Engenharia Industrial.

FISWG *Facial Identification Scientific Working Group.*

IQR Intervalo Interquartil.

ISV *inter-session variability.*

KDE *Kernel Density Estimation.*

LBPH *Local Binary Pattern Histogram*

LDA *Linear Discriminant Analisys.*

LFW *Labeled Faces in the wild.*

LR *Likelihood-Ratio.*

MTCNN *Multi-Task Cascaded Convolutional Neural Networks.*

NBE *Normalized Bayes Error-rate.*

NIFS/ANZPAA *National Institute of Forensic Science Australia New Zealand.*

OSAC *Organization of Scientific Area Committess for Forensic Science.*

PAV *Pool Adjacent Violators.*

PCA *Principal Component Analysis.*

PDF *Probability Density Function.*

PTZ *Pan, Tilt, Zoom.*

RL Regressão Logística

RLR Regressão Logística Regularizada.

RMEd *Rate of Misleading Evidence in favor of the defense.*

RMEp *Rate of Misleading Evidence in favor of the prosecution.*

ROC Receiver Operating Characteristic

ROCCH ROC convex hull

SCRFD *Sample and Computation Redistribution for Efficient Face Detection.*

SLR *Score-based* LR.

SS *Same Source.*

WSV *Within-Source Variability.*

Capítulo 1

Introdução

1.1 Motivação

A crescente disponibilidade de dispositivos como *smartphones* e câmeras de videomonitoramento tem tornado cada vez mais frequente que crimes dos mais variados tipos, como violência sexual, homicídios, tráfico de entorpecentes, tráfico de pessoas, entre outros, sejam registrados em imagens - fotografias ou vídeos [14].

Assim, além da possibilidade constatação do crime a partir de imagens, em muitos casos também é possível que a identificação do criminoso seja realizada a partir desse tipo de material. Para a identificação de pessoas a partir de imagens diversas abordagens são possíveis, a depender de quais características da pessoa de interesse estão disponíveis nas imagens, sendo o foco desta pesquisa a utilização de características faciais, ou seja, o exame pericial de comparação facial.

As principais recomendações técnicas e diretrizes para a realização deste tipo de exame são elaboradas pelo Grupo de Trabalho Científico em Identificação Facial (*Facial Identification Scientific Working Group* - FISWG), entidade que inclui representantes da indústria, de órgãos de pesquisa e de instituições policiais e periciais de países diversos. Atualmente o FISWG recomenda a análise e comparação morfológica como principal método a para comparação facial forense [15], sendo esse um processo manual, realizado por especialistas.

Além do FISWG, a Rede Europeia de Institutos de Ciências Forenses (*European Network of Forensic Science Institutes* - ENFSI) também recomenda a análise e comparação de características morfológicas como método de referência para este exame [16].

Um aspecto importante sobre este método é que os resultados são sempre expressos de forma qualitativa, variando entre conclusões categóricas (ex. “As imagens são de um mesmo indivíduo.”), conclusões em uma escala qualitativa de probabilidades *a posteriori* (ex. “É muito provável que as imagens sejam de um mesmo indivíduo.”) e conclusões

baseadas em uma escala qualitativa de razão de verossimilhança (do inglês *Likelihood Ratio* - LR) para descrever o peso da evidência obtida (ex. "As semelhanças e diferenças encontradas nas imagens comparadas são muito mais plausíveis considerando que as imagens têm como fonte um mesmo indivíduo do que considerando que elas têm como fonte indivíduos distintos da população de referência.").

A utilização do paradigma da LR para apresentação de resultados de exames periciais é recomendada por instituições como a ENFSI [17, 16] e o *National Institute of Forensic Science Australia New Zealand* (NIFS/ANZPAA) [18]. Além disso, encontra-se em elaboração pela *Organization of Scientific Area Committess for Forensic Science* (OSAC), entidade norte-americana encarregada de coordenar grupos de trabalho nas áreas de ciências forenses, uma orientação[1] para que a conclusão de exames de comparações por imagens sejam expressas através de uma escala que leve em conta, necessariamente, duas hipóteses (ou proposições) mutuamente excludentes relacionadas à origem do vestígio, se aproximando de uma abordagem baseada em LR.

Especificamente em relação às conclusões do exame de comparação facial dentro de uma abordagem baseada em LR, tal conclusão apresenta uma descrição qualitativa do nível de suporte relativo de uma proposição em relação à outra [19]. A determinação pelo especialista do nível de suporte é uma tarefa complexa e que depende de fatores inter-relacionados, como os seguintes:

- quantidade de características faciais que estão visíveis nas imagens;
- quantidade de características similares e dissimilares nas imagens comparadas;
- o quão raras são as características observadas nas imagens (baseado na experiência e conhecimento subjetivo do especialista);
- a quão transitórias ou permanentes são as características observadas;
- se as similaridades e dissimilaridades podem ser explicadas por fatores relacionados à qualidade das imagens;
- a correspondência entre as condições de aquisição de cada imagem comparada (ex. iluminação, enquadramento, resolução espacial);
- a diferença de tempo entre a captura das imagens, entre outros.

[1] OSAC 2022-S-0001, *Standard Guide for Image Based Comparison Conclusions/Opinions*, versão 1.0 proposta, disponível em `https://www.nist.gov/document/osac-2022-s-0001-standard-guide-image-comparison-conclusions-opinionsfor-open-comment2`, acesso em 16/05/2023.

A consideração em conjunto de todos esses fatores é uma dificuldade importante para que peritos distintos obtenham o mesmo resultado em determinado exame [20]. Alguns fatores dependem essencialmente da experiência e de conhecimentos subjetivos de cada perito, enquanto para outros não existem levantamentos estatísticos que possam embasar avaliações quantitativas, por exemplo, sobre o grau de transitoriedade de determinada característica morfológica.

Cabe observar, porém, que apesar dessas dificuldades, peritos com experiência e treinamento para o exame já foram testados e foi verificado que apresentam acurácia superior a de pessoas não treinadas [21, 22].

A apresentação dos resultados de forma qualitativa, ainda que dentro do paradigma de LR, possui algumas limitações, como a já mencionada possibilidade de que peritos distintos poderiam, avaliando as mesmas imagens e obtendo o mesmo conjunto de semelhanças e diferenças, escolherem níveis diferentes da escala para a conclusão. Outra limitação se refere à dificuldade de combinação do resultado qualitativo com os resultados de outros exames periciais relevantes no caso [23].

Além de permitirem expressar os resultados apenas de forma qualitativa, os métodos atuais não incluem o uso de sistemas biométricos no exame, apesar do desempenho de sistemas de reconhecimento facial terem ultrapassado o desempenho de humanos em alguns cenários desde 2015 [24, 25, 26].

É de se destacar que sistemas biométricos de reconhecimento facial tiveram um aumento substancial de desempenho nos últimos anos [26, 25], especialmente devido à disponibilidade de grandes bases de dados para treinamento e de arquiteturas de redes neurais convolucionais profundas, como as ResNets [27].

Além da evolução no desempenho desses sistemas, a utilização conjunta de *experts*, empregando análises manuais tradicionais, com sistemas automáticos de reconhecimento facial apresenta melhor acurácia em tarefas de verificação (comparação 1:1) do que apenas *experts* ou apenas sistemas biométricos, isoladamente. A Figura 1.1, extraída de [1], ilustra esse resultado.

Para que esta utilização conjunta seja possível em cenários forenses, um primeiro passo consiste em desenvolver metodologia que permita a interpretação dos resultados de sistemas biométricos sob a forma de LR.

Assim, nesta pesquisa serão avaliados métodos que permitam expressar os resultados do exame de comparação facial de forma quantitativa, com o cálculo de LR, a partir da utilização de sistemas automáticos de reconhecimento facial.

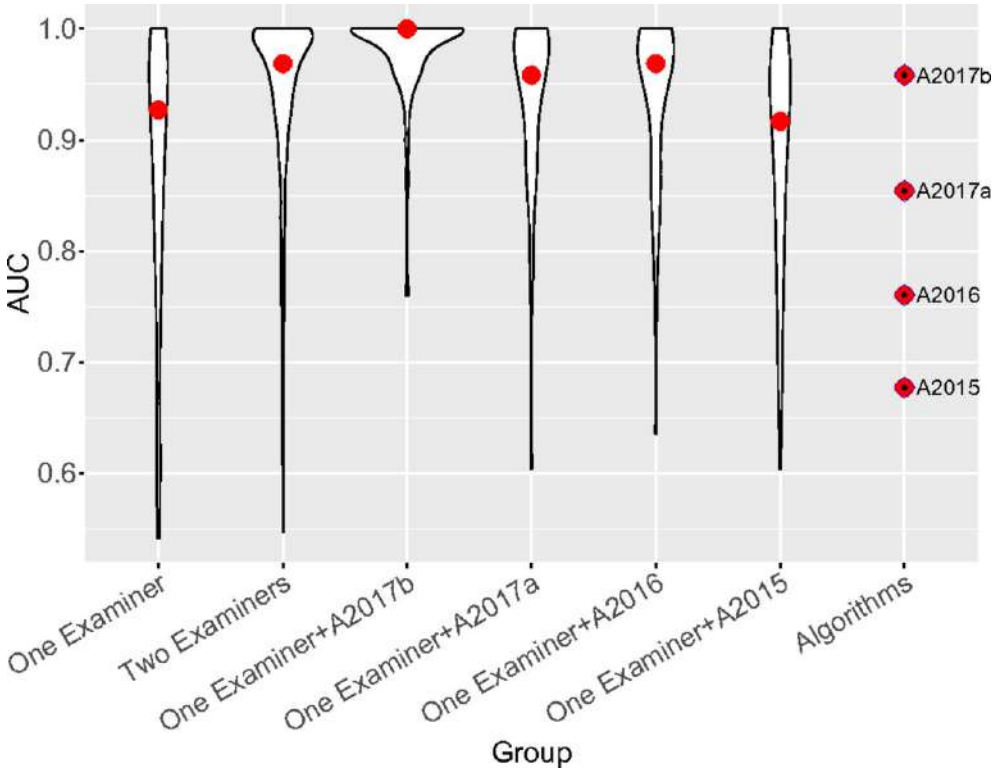


Figura 1.1: Desempenho de diferentes grupos para reconhecimento facial (1:1). O eixo vertical mostra a distribuição de desempenho em termos de AUC (Área sob a curva ROC) para cada grupo e o ponto vermelho mostra a mediana de cada grupo. A combinação de um perito com o melhor algoritmo testado resultou em AUC mediana de 1,0, resultado melhor do que combinando as respostas de dois peritos. Reproduzida de [1], com permissão.

1.2 Objetivos

1.2.1 Objetivo Geral

A presente pesquisa tem como objetivo avaliar métodos para emprego de sistemas de reconhecimento facial em exames periciais, com a expressão dos resultados do exame de forma quantitativa, através de LR.

1.2.2 Objetivos Específicos

1. Identificação, obtenção e caracterização de bases de imagens faciais representativas dos contextos mais comuns em perícias de comparação facial:

   Como será tratado no Capítulo 2, o cálculo de LR depende da disponibilidade de bases de imagens faciais representativas dos casos tratados nos exames periciais. Assim, a identificação, obtenção e caracterização dessas bases constituem elemento essencial a esta pesquisa.

2. Validação de métodos de cálculo de LR a partir de sistemas de reconhecimento facial:

   A presente pesquisa envolve a realização de experimentos que dependem da extensão de modelos de reconhecimento facial, com a adição de blocos funcionais destinados à obtenção de escores de similaridade, ao processo de conversão de escores em LR[2] e à verificação do desempenho (validação) dos modelos de cálculo de LR.

1.3 Organização da dissertação

Este trabalho é dividido da seguinte forma: no Capítulo 2 são apresentados conceitos teóricos relacionados à interpretação de evidência para fins forenses com foco no paradigma da LR, necessários à compreensão de temas que serão tratados em seguida, no Capítulo 3, no qual são apresentados trabalhos relacionados a sistemas de reconhecimento facial e ao emprego desses sistemas no contexto pericial. No Capítulo 4 a metodologia proposta, os dados utilizados e os critérios de validação são descritos. No Capítulo 5 são apresentados os resultados dos experimentos para validação de sistemas de cálculo de LR a partir de escores, incluindo experimentos relacionados à agregação de *embeddings* em cenários em que estejam disponíveis múltiplas imagens do mesmo indivíduo, e no Capítulo 6 são apresentadas as conclusões da pesquisa, as limitações encontradas e os trabalhos futuros.

[2]Alguns autores denominam o processo de conversão de escores em LR de calibração, enquanto outros utilizam o mesmo termo para se referir a uma propriedade de sistemas de cálculo de LR (o quão calibrado é um sistema). Neste trabalho é adotado o segundo sentido descrito.

Capítulo 2

Conceitos Teóricos

Neste capítulo são apresentados conceitos teóricos relacionados à interpretação de evidência para fins forenses segundo o paradigma da LR.

2.1 Interpretação de evidência sob o paradigma da LR

Este capítulo tem o objetivo de apresentar o paradigma da LR para interpretação de evidências em processos judiciais.

No caso desta pesquisa, trata-se da possibilidade de apresentar como evidência o resultado da comparação entre duas imagens faciais, realizada com auxílio de um sistema biométrico, de uma forma logicamente adequada à interpretação no sistema judicial e que permita a apreciação conjunta deste tipo de prova com as demais evidências no processo.

Assim, o que se busca é apresentar, como resultado da comparação feita por um sistema biométrico, o chamado peso da evidência.

O peso da evidência, considerando exames periciais de comparação facial, representa o quanto a evidência fortalece ou enfraquece a hipótese de as imagens representarem o mesmo indivíduo, em contraposição à hipótese de representarem indivíduos distintos. Essas duas hipóteses são geralmente chamadas de hipótese da acusação e de hipótese da defesa, respectivamente.

Este tipo de abordagem para interpretação e apresentação do peso da evidência, chamado de paradigma da LR (*likelihood ratio*, ou razão de verossimilhança), é baseado no modelo *Case Assessment and Interpretation*, formulado em [28, 29]. Segundo esse paradigma, a interpretação da evidência é baseada no cálculo da LR, que depende da avaliação da probabilidade de se obter a evidência considerando duas hipóteses mutuamente excludentes. A LR é calculada conforme a seguinte expressão:

$$LR = \frac{Prob.\ da\ evidência\ se\ a\ hipótese\ da\ acusação\ é\ verdadeira}{Prob.\ da\ evidência\ se\ a\ hipótese\ da\ defesa\ é\ verdadeira} = \frac{P(E|H_p, I)}{P(E|H_d, I)}. \quad (2.1)$$

No termo à direita da Equação 2.1, E representa a evidência, H_p e H_d representam as hipóteses de acusação e defesa, respectivamente, e I significa as informações de contexto relevantes para a apreciação da evidência. Assim, $P(E|H_p, I)$ representa a probabilidade de obter a evidência E, considerando que a hipótese da acusação é verdadeira e levando em conta as informações de contexto relevantes ao exame. $P(E|H_d, I)$, por sua vez, representa a probabilidade de obter a mesma evidência E, mas considerando verdadeira a hipótese da defesa e as mesmas informações de contexto.

A LR é um dos termos do Teorema de Bayes sob a forma de razões de probabilidades, que pode ser descrito pela seguinte equação, utilizando a notação já apresentada:

$$\underbrace{\frac{P(H_p|E, I)}{P(H_d|E, I)}}_{\text{Razão de prob. } a\ posteriori} = \underbrace{\frac{P(E|H_p, I)}{P(E|H_d, I)}}_{\text{LR}} \times \underbrace{\frac{P(H_p|I)}{P(H_d|I)}}_{\text{Razão de prob. } a\ priori} \quad (2.2)$$

A Equação 2.2 pode ser compreendida como um mecanismo para atualização da convicção[1] da instância julgadora em relação às hipóteses de interesse (H_p e H_d), após a apresentação da evidência E e considerando as informações de contexto I.

O termo $\frac{P(H_p|I)}{P(H_d|I)}$ representa a razão entre a probabilidade da hipótese de acusação (ex. "as imagens são da mesma pessoa") e a probabilidade da hipótese da defesa (ex. "as imagens são de pessoas diferentes") *a priori*, ou seja, antes da apreciação da evidência e dadas as demais informações relevantes do caso (inclusive outras evidências previamente apresentadas).

O termo $\frac{P(E|H_p,I)}{P(E|H_d,I)}$ é a LR, ou seja, o resultado a ser apresentado pela perícia, que representa o quanto a evidência obtida é mais provável considerando verdadeira a hipótese da acusação do que considerando verdadeira a hipótese da defesa.

Por fim, o termo à esquerda, $\frac{P(H_p|E,I)}{P(H_d|E,I)}$, representa a razão entre as probabilidades das hipóteses da acusação e da defesa *a posteriori*, ou seja após a apreciação da evidência pelo julgador.

Além de representar um mecanismo para atualização da convicção da instância julgadora frente a novas evidências, o paradigma da LR naturalmente delimita os papéis da instância julgadora e da perícia. Ao cientista forense, ou à perícia, cabe examinar

[1]Por convicção da instância julgadora pode ser entendida a razão entre as probabilidades de interesse para o caso, ou seja, o quanto a hipótese da acusação é mais ou menos plausível do que a hipótese da defesa.

o vestígio, utilizando conhecimento especializado, e avaliar a plausibilidade de obter a evidência em questão considerando cada uma das hipóteses. Essa avaliação é sumarizada pela LR, a qual deve ser então empregada pela instância julgadora para atualizar a sua convicção em relação às hipóteses de interesse. Deve-se notar que a perícia, ao examinar um vestígio específico, usualmente não tem conhecimento a respeito das demais evidências do caso ou de outras informações relevantes que devem ser levadas em consideração no processo decisório. Assim, tanto a razão de probabilidade das hipóteses *a priori* quanto *a posteriori* são consideradas como atribuições da instância julgadora. [17, 28, 29]

É fundamental notar que nesse paradigma não cabe à perícia tomar uma decisão a respeito da origem do vestígio, como é comum em aplicações biométricas através da utilização de um limiar de decisão [17]. Neste paradigma, o resultado a ser apresentado pela perícia, o peso da evidência, deve fornecer um grau de suporte para uma hipótese de origem do vestígio, relativamente à outra hipótese considerada. Tal distinção será retomada no Capítulo 4, pois serão necessárias métricas de performance diferentes daquelas utilizadas em sistemas biométricos clássicos para a avaliação de sistemas de cálculo de LR.

2.2 Hierarquia das proposições

Outro aspecto relevante no paradigma da LR é o conceito de hierarquia das proposições, definido em [29]. Essa hierarquia define três níveis de proposições (ou de hipóteses):

I. nível do crime;

II. nível da atividade; e

III. nível da fonte.

São consideradas proposições do primeiro nível aquelas que dizem respeito diretamente ao crime em apuração (por exemplo, "Fulano de tal cometeu o assassinato" *versus* "Outra pessoa cometeu o assassinato"). Evidentemente, proposições deste tipo são as de maior relevância para a formação de decisão por parte do julgador mas, por outro lado, envolvem avaliações que muito frequentemente estão fora do domínio de conhecimento e atribuição da perícia (por exemplo, quanto a eventual alegação de legítima defesa e enquadramento legal de determinada conduta).

O nível da atividade engloba hipóteses que se referem a ações que resultaram no vestígio analisado. Por exemplo, o par de proposições "Fulano de tal atirou em Beltrano" *versus* "Outra pessoa atirou em Beltrano" seriam consideradas como proposições deste nível. Neste nível de proposição, com alguma frequência, é possível a avaliação da evidência

pelo cientista forense, embora ela seja mais complexa do que a avaliação de proposições do nível da fonte e frequentemente requer informações não disponíveis à perícia.

No nível da fonte, as proposições se limitam às possíveis origens do vestígio examinado. Por exemplo "O projétil de munição coletado no local do crime partiu da arma de Fulano de Tal" *versus* "O projétil foi disparado por outra arma de mesmo calibre" seriam consideradas proposições do nível da fonte.

Proposições no nível da fonte estão mais distantes do que, em última análise, interessa à instância julgadora, ou seja, a culpa ou inocência do suspeito. Entretanto, proposições no nível da fonte geralmente são aquelas para as quais os cientistas forenses estão mais aptos a oferecerem respostas.

No caso das hipóteses consideradas nesta pesquisa, elas se enquadram neste nível (da fonte), pois usualmente são elaboradas nos seguintes termos:

> H_p = *As imagens faciais examinadas são de um mesmo indivíduo.*
>
> H_d = *As imagens faciais examinadas são de indivíduos diferentes da população de referência.*

É usual, portanto, que exames periciais que avaliam hipóteses desse nível sejam denominados de exames periciais de atribuição de fonte.

2.3 Evidência e Hipóteses

2.3.1 Evidência

No contexto de exames periciais de atribuição de fonte e considerando o paradigma da LR, a **Evidência** deve ser compreendida como o conjunto de similaridades e diferenças obtido da comparação entre as características do vestígio (por exemplo, a imagem do autor de um crime) e do suspeito (por exemplo, imagens faciais de prontuários de identificação ou de documentos de identidade do suspeito).

No caso da utilização de sistemas biométricos para a fins forenses, como nesta pesquisa, a evidência a ser considerada é o escore obtido pelo sistema ao se comparar a amostra biométrica do vestígio com a amostra biométrica do suspeito [7, 30, 8, 31].

2.3.2 Hipóteses

No paradigma da LR, as hipóteses a serem consideradas e contrapostas quando da avaliação do peso da evidência devem ser mutuamente excludentes, embora não seja necessário que elas sejam complementares, ou seja, não é requerido que

$$P(H_p|I) + P(H_d|I) = 1. \tag{2.3}$$

Um aspecto importante em relação às hipóteses no nível da fonte se refere às hipóteses serem ou não ancoradas em um suspeito específico.

Uma hipótese como "a imagem questionada tem como fonte o suspeito Fulano de Tal" é um exemplo de hipótese ancorada no suspeito, ao passo que "as imagens sob exame são de uma mesma pessoa" é uma hipótese não-ancorada.

Deve-se destacar que hipóteses ancoradas no suspeito apresentam uma dificuldade adicional, pois é necessário obter várias imagens do suspeito, em condições semelhantes às imagens do caso, para que a modelagem estatística de cada hipótese seja robusta.

Por esta razão, nesta pesquisa as hipóteses consideradas são sempre do tipo não-ancoradas, ou seja, referem-se apenas às imagens terem como origem um mesmo indivíduo ou terem como origens indivíduos distintos da população de referência. No paradigma da LR e considerando exames de atribuição de fonte, população de referência é a população de potenciais fontes alternativas do vestígio.

2.4 Tipos de escores

Conforme mencionado na Subseção 2.3.1, a evidência a ser considerada para o cálculo de LR a partir de sistemas de reconhecimento facial é o escore obtido na comparação entre as imagens sob exame.

É necessário fazer uma sucinta discussão sobre dois tipos de escores que podem ser utilizados para essa finalidade, a saber:

- escores de similaridade
- escores de similaridade e tipicidade

Escores de similaridade consideram apenas a proximidade (ou distância) entre dois objetos, sem levar em consideração o quanto cada objeto é típico na população de interesse.

Por outro lado, escores de similaridade e tipicidade consideram também esse segundo aspecto, o que sugere que este tipo de escore é mais apropriado para o cálculo de LR do que escores que consideram apenas a similaridade entre as amostras comparadas.

De fato, [32] demonstrou, através de simulações Monte-Carlo, a superioridade de escores que consideram tanto a similaridade quanto a tipicidade, comparativamente a escores que consideram apenas similaridade para a obtenção de LR.

Nesta pesquisa, serão realizados experimentos apenas com escores de similaridade. Experimentos com escores de similaridade e tipicidade serão realizados em trabalho futuro.

Capítulo 3

Trabalhos Relacionados

Este capítulo oferece uma revisão da literatura relacionada a reconhecimento facial e ao emprego de sistemas biométricos para fins forenses. São também apresentadas bases de imagens faciais relevantes para esta pesquisa.

3.1 Sistemas de Reconhecimento Facial

A utilização de computadores para reconhecer faces remonta à década de 1960, quando os trabalhos [33, 34, 35, 36] apresentaram um sistema em que coordenadas de pontos da face eram obtidas manualmente, marcadas por operadores humanos em um dispositivo específico. Dessas coordenadas era obtida uma lista de 20 distâncias, como larguras da boca e dos olhos, distâncias inter-pupilares, entre outras, que eram então processadas por computador para efetuar o reconhecimento, comparando as distâncias obtidas com aquelas correspondentes a outras pessoas previamente registradas na base. Deve-se destacar, porém, que esta abordagem não envolve visão computacional, uma vez que a visualização e localização das características utilizadas para reconhecimento era feita por operadores humanos.

Em 1973, [37] apresentou o primeiro sistema de reconhecimento facial completamente automático, no qual as características utilizadas para reconhecimento eram obtidas pelo sistema utilizando técnicas de visão computacional. A extração de características era precedida por um estágio de pré-processamento para detecção de linhas de contorno utilizando um operador laplaciano. A imagem pré-processada era então submetida a sub-rotinas diversas, cada uma projetada para a detecção de elementos específicos da face, como olhos, nariz, boca e contornos da face. Além disso, havia um esquema de realimentação entre as sub-rotinas, que tinha como objetivo manter a consistência entre os resultados de cada sub-rotina - por exemplo, para evitar resultados como a detecção da boca acima do nariz ou fora dos contornos da face. Pontos específicos eram então obti-

dos e relações de distâncias e ângulos entre pontos eram utilizadas para comparação. A Figura 3.1 ilustra um exemplo de uma imagem submetida ao sistema e os resultados das etapas de pré-processamento, localização dos pontos e as relações de distâncias e ângulos consideradas para comparações.

Figura 3.1: Exemplo de resultados das etapas do sistema de T. Kanade (três imagens ao centro) e distâncias e ângulos considerados (à direita). Figura extraída e adaptada de [2].

O sistema foi utilizado para localizar pontos e ângulos de 853 imagens faciais e processou corretamente 673 delas. Considerando apenas imagens com características para as quais o sistema foi originalmente projetado, ou seja, imagens frontais e sem presença de barba ou óculos, 608 imagens foram corretamente processadas, de um total de 670. O sistema também foi testado para identificação em um conjunto de 20 pessoas, identificando corretamente 15 delas (75%), o mesmo desempenho obtido ao se utilizar pontos e ângulos marcados por humanos.

A próxima evolução significativa em sistemas de reconhecimento facial ocorreu a partir do final da década de 1980, quando [38] propôs a utilização de análise por componentes principais (do inglês *Principal Component Analysis* - PCA) para representar imagens faciais em um espaço de dimensionalidade reduzida.

Em 1991, [39] aplicou esta técnica e propôs um sistema de quase-tempo real capaz de localizar e rastrear faces em vídeos e também reconhecê-las, comparando-as com faces de indivíduos conhecidos pelo sistema. Nesta abordagem, cada imagem facial é representada através de sua projeção em um espaço composto pelas componentes que representam as maiores variações das faces conhecidas pelo sistema. Estas componentes são chamadas de *eigenfaces*, por serem os autovetores (*eigenvectors*) do conjunto de faces utilizadas na etapa de treinamento. A operação de projeção caracteriza cada face como uma soma ponderada das *eigenfaces*, de modo que o reconhecimento é feito comparando-se os pesos correspondentes à projeção de cada face com os pesos das demais faces conhecidas pelo sistema.

As etapas necessárias para o treinamento do sistema proposto por [39] eram as seguintes:

1. Obter um conjunto de imagens faciais para treinamento;

2. Calcular as *eigenfaces* do conjunto de treinamento, mantendo apenas as M imagens que correspondem aos maiores autovalores. Estas M imagens definem o *espaço de faces*. À medida em que novas faces são apresentadas ao sistema, o espaço de faces pode ser atualizado; e

3. Projetar as imagens faciais do conjunto de treinamento no espaço de faces, através da obtenção dos pesos necessários para representar cada imagem como uma soma ponderada das *eigenfaces*.

Após a etapa de treinamento, o reconhecimento de uma imagem apresentada ao sistema era realizado da seguinte maneira:

1. Calcular o conjunto de pesos que representa a imagem no espaço de faces, projetando a imagem em cada uma das *eigenfaces*;

2. Determinar se a imagem representa uma face ou não, calculando a distância entre sua projeção e o espaço de faces;

3. Se for uma face, comparar com as faces conhecidas pelo sistema para fazer o reconhecimento;

4. (Opcional) Atualizar o espaço de faces;

5. (Opcional) Se a mesma face desconhecida é apresentada várias vezes, obter a média das representações das imagens daquela face e incluí-la como uma nova identidade no sistema.

Foram relatados dois experimentos pelos autores, nos quais se verificou que o sistema era relativamente robusto a variações de iluminação, menos robusto a variações de orientação (rotação da face na imagem) e era significativamente afetado por variações de resolução.

Já em 1997, [40] propôs o uso de análise por discriminantes lineares (do inglês *Linear Discriminant Analisys* - LDA) para reconhecimento facial. Esta técnica é denominada também como *fisherfaces*, em referência a Robert Fisher, formulador da LDA [41]. Neste caso, a transformação das imagens faciais é realizada levando-se em conta a maximização da separação entre classes (identidades) e a minimização das distâncias intra-classes, enquanto no caso da abordagem por *eigenfaces*, a transformação buscava representações que capturassem a maior variabilidade do conjunto de treinamento, sem considerar as separações entre classes.

As abordagens *eigenfaces*, *fisherfaces* e suas derivadas podem ser classificadas como holísticas, pois buscam uma representação global da imagem facial. Esse tipo de abordagem apresenta dificuldades em reconhecer imagens com variações em pose, expressão facial, iluminação, entre outros. Assim, técnicas baseadas em descritores locais de elementos da face foram desenvolvidas a partir do início dos anos 2000. Gabor [42] e Histogramas de Padrões Binários Locais (do inglês *Local Binary Pattern Histogram* - LBPH) [43] são dois exemplos proeminentes deste tipo de técnica em que se tentava projetar descritores locais para reconhecimento facial, o que pode ser interpretado como uma retomada das ideias expostas em [2].

Já no início da década de 2010, foram apresentadas abordagens que utilizam inteligência artificial para o aprendizado de descritores locais [44, 45]. Estas técnicas, porém, ainda não utilizavam aprendizado profundo.

Em 2014, [46] apresentou técnica de aprendizado profundo para o aprendizado de descritores. No mesmo ano, [3] e [47] empregaram redes neurais convolucionais profundas para reconhecimento facial, alcançando desempenho equivalente ao de humanos, considerando a base LFW [48]. Alguns autores [49] indicam a possibilidade de que o desempenho de humanos para reconhecimento de faces não familiares tomando a base LFW como referência esteja fortemente superestimada, uma vez que as faces presentes nessa base são, em sua maioria, de pessoas famosas. Além disso, as avaliações de desempenho de seres humanos geralmente utilizam algum processo de fusão de escores obtidos de diversas pessoas, o que tende a melhorar o desempenho comparativamente à média do desempenho das pessoas.

Embora não seja o objetivo desta pesquisa estudar sistemas biométricos de reconhecimento facial, é oportuno tratar brevemente da arquitetura do modelo proposto por [3], baseado em redes neurais convolucionais, uma vez que este tipo de rede é até hoje utilizado nos modelos com maior desempenho, inclusive nos modelos utilizados nesta pesquisa.

A arquitetura proposta em [3], conhecida como DeepFace e ilustrada na Figura 3.2, é composta por nove camadas. A primeira camada é responsável pelo alinhamento da imagem da face, que envolve a frontalização a partir de modelagem tridimensional da face. Há em seguida uma sequência de três camadas, uma de convolução, uma de *pooling* máximo e outra de convolução, com o objetivo de obter descritores de baixo nível, como bordas e texturas. As três camadas seguintes são camadas localmente conectadas que, diferentemente das camadas convolucionais, não possuem seus pesos compartilhados. A camada seguinte é do tipo totalmente conectada e é capaz de capturar correlações entre elementos distantes na imagem. Esta camada também é aquela em que se obtém uma representação da face, com 4.096 elementos, utilizada como entrada para a última camada, também totalmente conectada, em que cada nó representa uma classe da base utilizada

no treinamento da rede.

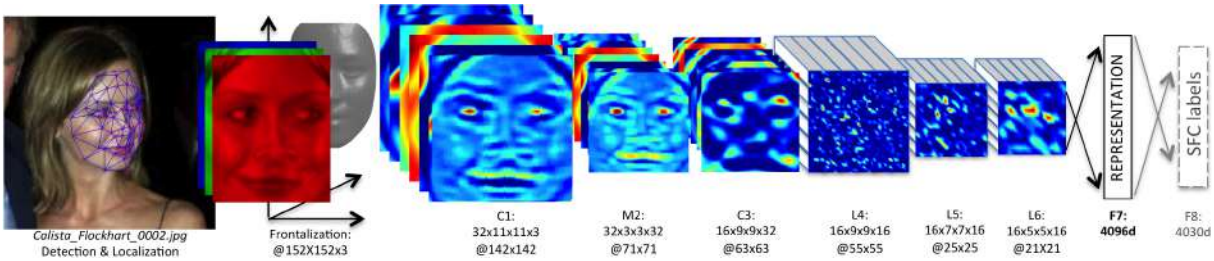


Figura 3.2: Arquitetura do DeepFace. Reproduzida de [3], com permissão. © 2014 IEEE.

Uma característica relevante em redes neurais convolucionais utilizadas para reconhecimento facial (e para outras finalidades de processamento de imagens) é que as características utilizadas para classificação são aprendidas pela rede, ao invés de serem projetadas previamente, como nas abordagens clássicas. Assim, é interessante tentar compreender quais características são aprendidas por este tipo de rede durante o treinamento. [4] investigou esta questão e demonstrou que, em geral, as camadas inicias de redes convolucionais aprendem características relacionadas a formas mais básicas, como linhas verticais, horizontais e diagonais. As camadas subsequentes aprendem características cuja complexidade e significado tendem a aumentar conforme se avaliam camadas mais profundas. A Figura 3.3 ilustra um mapa de características de uma rede AlexNet [50].

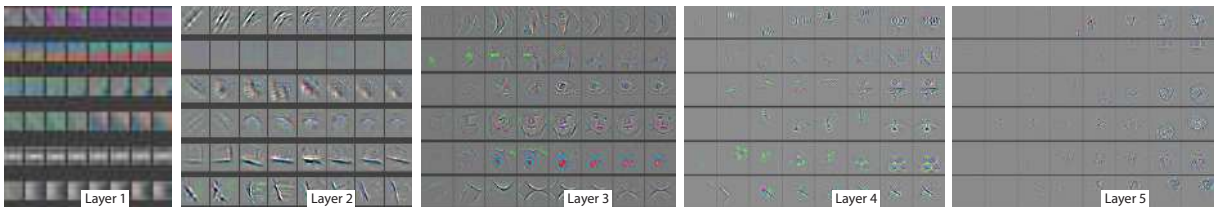


Figura 3.3: Mapas de características de camadas convolucionais de uma rede AlexNet. Reproduzida de [4], com permissão.

Em relação ao modelo proposto por [47], embora sua arquitetura seja bastante semelhante à do DeepFace, o treinamento do modelo teve um aspecto inovador, pois considerou dois objetivos relacionados ao reconhecimento facial: a identificação, que consiste em atribuir à imagem uma classe dentre todas as presentes no conjunto de treinamento; e a verificação, que consiste em verificar se duas imagens foram obtidas de uma mesma pessoa.

Em 2015 um novo marco de desempenho foi alcançado com o modelo FaceNet [24], que obteve 99.63% de acurácia na base LFW [48]. Este modelo emprega um estágio de pré-processamento mais simples do que o DeepFace, apenas com alinhamento em duas dimensões. Uma contribuição muito relevante deste trabalho é a introdução de uma nova função de perda para treinamento da rede neural, chamada de *triplet loss*, em que

são submetidas três imagens: uma chamada de âncora, outra que forma um par positivo (mesma identidade da âncora) e outra que forma um par negativo (identidade diferente da âncora). A rede é treinada para minimizar a distância entre as imagens do par positivo e aumentar a distância entre aquelas do par negativo. É possível identificar uma semelhança de objetivos com a abordagem LDA.

A esta altura, as pesquisas na área de reconhecimento facial passaram a se concentrar em três aspectos principais: arquitetura da rede, bases de treinamento e funções de perda.

De fato, além de apresentar uma nova função de perda, o trabalho [24] explorou diferentes arquiteturas (Inception[51] e Zeiler&Fergus [52]) e o impacto da quantidade de imagens de treinamento (entre 2,6 milhões e 260 milhões de imagens). Foi demonstrado o impacto positivo de arquiteturas mais profundas (com mais camadas) e de bases de treinamento com mais imagens.

Em 2016 foi apresentada uma nova arquitetura de rede, chamada ResNet [27]. A principal inovação desta arquitetura é a introdução de *skip connections*, ou seja, de conexões diretas entre camadas não consecutivas. Isso viabilizou a construção e treinamento de modelos com dezenas ou até mesmo centenas de camadas, o que se demonstrou extremamente eficaz em tarefas como reconhecimento e classificação de imagens.

Modelos para detecção de faces e para reconhecimento facial baseados em arquitetura ResNet e derivadas estão, ainda hoje, entre aqueles com os melhores desempenhos em *benchmarks* como MegaFace [53] (para reconhecimento facial [54]) e WIDER Face Hard [55] (para detecção de faces [56]).

Se, por um lado, a utilização de arquiteturas extremamente profundas possibilitou alcançar desempenhos muito elevados, a complexidade e os requisitos de memória e processamento se colocam como um impeditivo prático para o uso desses modelos em alguns cenários. Assim, algumas arquiteturas que privilegiam o baixo consumo de memória e processamento foram propostas nos últimos anos, com destaque para MobileNet v1 [57] e v2 [58], ShuffleNet [59] e EfficientNet [60].

3.2 Bases de imagens faciais

A evolução nas arquiteturas de redes neurais profundas foi acompanhada pela disponibilidade de bases de imagens faciais para treinamento cada vez maiores. A Tabela 3.1 relaciona algumas dessas bases, organizadas cronologicamente.

Destaca-se que as bases relacionadas na Tabela 3.1 são utilizadas, primariamente, para o treinamento de modelos de reconhecimento facial, sendo, em geral, constituídas por imagens capturadas em condições variadas, o que permite que os modelos treinados a

Base	Ano	n. de imagens (mil)	n. de identidades (mil)
LFW[48]	2007	13	5,7
CASIA-WebFace[61]	2014	500	10,5
VGGFace[62]	2015	2.600	2,6
MegaFace[53]	2016	1000	690
MS-Celeb-1M[63]	2016	10.000	100
Glint360K[64]	2020	17.000	360

Tabela 3.1: Bases de imagens faciais

partir dessas bases possuam desempenho satisfatório em uma variedade de condições de pose, iluminação, expressões faciais, resolução, entre outros.

Por outro lado, essas bases apresentam utilidade limitada para o desenvolvimento e validação de modelos de conversão de escore de similaridade para LR, ou seja, para o desenvolvimento de sistemas de cálculo de LR para fins forenses. Para este tipo de sistema, é necessário o emprego de conjuntos de imagens representativas dos casos apresentados à perícia. Como exemplo, se um determinado exame pericial recai sobre imagens capturadas em pose frontal, com iluminação uniforme e expressão facial neutra, é preciso que o sistema de cálculo de LR empregado neste exame tenha sido obtido a partir de imagens em condições semelhantes.

Assim, para este trabalho, foram selecionadas algumas bases, descritas na Tabela 3.2 que possuem características que refletem a casuística das perícias de comparação facial, que usualmente recai sobre imagens de documentos de identificação, com captura em condições controladas, sobre imagens de sistemas de CFTV - circuito fechado de televisão, com captura não controlada e fatores como borrão de movimento, baixa resolução efetiva e baixo contraste na região facial. Mais recentemente imagens *in the wild* provenientes de redes sociais também passaram a constituir uma porção importante da casuística pericial.

Base	Ano	n. de imagens	n. de identidades
FEI [65]	2006	2.800	199
SCFace [5]	2011	4.160	130
Quis-Campi [66]	2017	3.000	320
ForenFace [67]	2017	2.819	97
Adience [68]	2014	26.580	2.284
BFW [69]	2020	20.000	800

Tabela 3.2: Bases de imagens faciais representativas de cenários forenses.

A base FEI foi constituída entre 2005 e 2006 no Laboratório de Inteligência Artificial da

FEI e é constituída por 2.800 imagens, sendo 14 imagens de cada um dos 199 indivíduos representados. Embora inicialmente projetada para conter imagens de 200 indivíduos, foi confirmado com o mantenedor da base que dois indivíduos supostamente diferentes, identificados na base como 002 e 071, são, na verdade, a mesma pessoa. Por isso, há 28 imagens deste indivíduo. As imagens foram capturadas em condições controladas de iluminação, pose e expressão facial, com fundo uniforme. Em relação à divisão por gênero, a base é equilibrada, sendo 100 mulheres e 99 homens, enquanto a distribuição por idade se concentra na faixa de 19 a 40 anos. A Figura 3.4 mostra exemplos das imagens existentes na base.

Figura 3.4: Exemplos de imagens presentes na base FEI.

A base FEI, embora não apresente desafios significativos para o desempenho dos sistemas atuais de reconhecimento facial, é relevante para o trabalho por permitir a validação do cálculo de LR considerando três aspectos: i) imagens de ótima qualidade; ii) utilização de imagens representativas de um subgrupo da população brasileira (especialmente adultos jovens); e iii) utilização de imagens representativas do contexto de perícias em imagens de documentos ou bases de dados de identificação.

A base SCFace foi publicada em 2011 e é composta por 4.160 imagens de 130 pessoas. Além de imagens em posição frontal capturadas com câmera fotográfica, foram capturadas imagens por cinco câmeras de CFTV, em ambiente interno e com iluminação não controlada, em três distâncias distintas - 1,0 m, 2,6 m e 4,2 m. Há também imagens capturadas por câmeras de CFTV em modo infra-vermelho. A Figura 3.5 ilustra algumas das imagens disponíveis na base.

A base SCFace permitirá explorar cenários típicos de exames periciais em que a imagem questionada, que representa o indivíduo de identidade desconhecida, é capturada em baixa resolução, por câmera de CFTV, e a imagem padrão, de um suspeito cuja identidade é conhecida, possui boa ou ótima qualidade, frequentemente podendo ter sido capturada pela própria equipe pericial com o propósito de ser utilizada como referência nos exames.

A base Quis-Campi é constituída por imagens de 320 pessoas capturadas por um conjunto de duas câmeras de CFTV - uma grande angular fixa e uma PTZ (*Pan, Tilt, Zoom*). Possui também vídeos e imagens estáticas capturados em condições variadas, in-

Figura 3.5: Exemplos de imagens presentes na base SCFace. À esquerda, imagem capturada por câmera fotográfica. À direita, imagens do mesmo indivíduo capturadas por cinco câmeras distintas de CFTV, a três distâncias. Adaptada de [5].

cluindo deslocamentos em ambientes externos e internos, com caminhada se aproximando e se afastando da câmera. Devido às imagens oriundas de câmeras de CFTV terem sido capturadas com os indivíduos em movimento, esta base apresenta imagens ainda mais semelhantes àquelas normalmente examinadas na perícia criminal do que a imagens contidas na base SCFace, por exemplo.

A base ForenFace também contém sequências de vídeo e imagens estáticas de câmeras de CFTV, capturadas de 97 pessoas. Foram empregadas seis câmeras de CFTV com configurações diversas, além de fotografias de alta qualidade de todos os indivíduos e nuvens de pontos capturadas com *scanner* 3D. O acesso à base ForenFace foi negado sob justificativa da incidência de novas regras relacionadas à privacidade das pessoas representadas na base, não existentes à época em que a base foi criada.

Por fim, a base BFW (do inglês *Balanced Faces in the Wild*) foi introduzida em 2020 tendo como premissa ser uma base balanceada em termos de gênero e grupos étnicos. As imagens são do tipo *in the wild*, apresentando maior dificuldade de reconhecimento do que a base LFW. Assim, a utilização da base BFW permitirá obter uma avaliação adequada a casos periciais em que as imagens provenham de redes sociais e apresentem baixa qualidade.

3.3 Agregação de *embeddings*

Em situações onde estão disponíveis múltiplas imagens faciais de um mesmo indivíduo, existe a possibilidade de agregar as informações das diferentes imagens a fim de obter uma representação mais fidedigna da identidade de cada pessoa. Em exames periciais é frequente que estejam disponíveis múltiplas imagens do indivíduo de interesse tanto no material padrão quanto no material questionado, especialmente em casos de vídeos. Uma abordagem que tem se demonstrado eficaz neste tipo de situação é a agregação das *embeddings* obtidas do modelo de reconhecimento facial [70, 71, 72].

Interessante notar que já no trabalho seminal de Turk e Pentland em 1991 [39] uma estratégia de agregação da representação de cada imagem facial era utilizada para representar uma nova identidade a ser adicionada ao sistema. A representação da identidade adicionada ao sistema era obtida a partir da média de cada componente das representações de cada imagem do mesmo indivíduo.

Em [70], é proposta uma estratégia de agregação simples das *embeddings* obtidas de múltiplas imagens do mesmo indivíduo para reconhecimento entre imagens estáticas e vídeo. A estratégia é baseada na média aritmética das componentes das *embeddings*. A mesma estratégia de agregação de *embeddings* é empregada em [71].

Outros trabalhos exploram a ideia de que imagens de melhor qualidade deveriam ter um peso maior no processo de agregação. [73] utilizou os escores de detecção facial como estimativa para qualidade da face e ponderou a agregação das *embeddings* por esta estimativa. [74, 75] empregam módulos específicos para agregação de *embeddings*. Os módulos são baseados em redes neurais e aprendem uma ponderação otimizada das *embeddings* que leva em conta a qualidade das imagens. Em [76], a ponderação por qualidade é tornada ainda mais específica, com o aprendizado de quais componentes das *embeddings* recebem maior peso no processo de agregação.

3.4 Sistemas de reconhecimento facial para fins forenses

A utilização de sistemas biométricos para fins forenses, com o cálculo de LR, tem se desenvolvido em ritmos diferentes entre as áreas especializadas de perícias. As áreas de DNA e reconhecimento de locutor, por exemplo, já utilizam sistemas biométricos com esta finalidade há bastante tempo [77, 78], enquanto na área de reconhecimento facial a adoção deste tipo de ferramenta é ainda incipiente.[8, 79, 30]

Em 2005, [6] propôs a utilização de uma abordagem baseada em análise bayesiana para interpretar evidências de impressão digital, face a assinatura, reforçando o caráter não-

específico da abordagem em relação à modalidade biométrica. Nesta abordagem sistemas biométricos eram utilizados para obter um escore (E) entre uma amostra questionada (vestígio) e uma amostra de referência. O mesmo sistema era empregado para obter dois conjuntos de escores a partir de uma população de referência. Esses conjuntos de escores eram utilizados para modelar duas funções densidade de probabilidade (PDF - *Probability Density Function*), uma relacionada à hipótese da acusação H_p (W), e a outra relacionada à hipótese da defesa H_d (B). O peso da evidência para o caso, ou seja, a LR, era obtido dividindo-se o valor da PDF que modela H_p (N) pelo valor da PDF que modela H_d (D), ambas avaliadas no valor do escore E do caso em análise. A Figura 3.6 ilustra a abordagem.

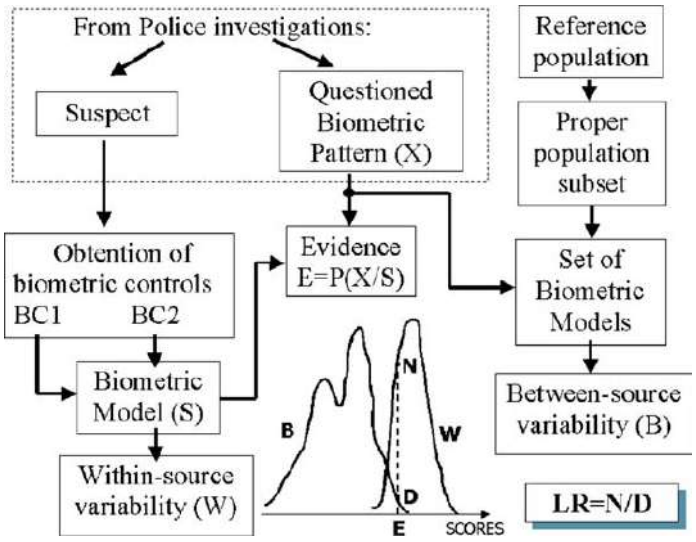


Figura 3.6: Cálculo de LR a partir de sistemas biométricos, conforme [6]. Reproduzida do original, com permissão.

Um aspecto importante nessa abordagem é de que a hipótese da acusação é ancorada no suspeito. Assim, a estimativa da função densidade de probabilidade W, que modela a intra-variabilidade da biometria do suspeito, deve ser feita a partir de escores obtidos entre uma amostra de referência do suspeito e de outras *amostras biométricas do suspeito coletadas em condições equivalentes às do vestígio*. Esta é uma dificuldade prática na maioria dos casos tratados pela perícia, pois geralmente é possível obter amostras de referência do suspeito (por exemplo, a partir de bases de dados de identificação ou de documentos oficiais), mas é dificílimo obter amostras do suspeito em condições equivalentes às do vestígio [80]. Em muitos casos seria necessário, por exemplo, levar o suspeito até o local onde foram gravadas as imagens do crime, em horário aproximado e com outras condições ambientais equivalentes.

De fato, essa dificuldade fez com que os autores propusessem estratégias para evitar uma estimativa muito baixa para a intra-variabilidade, fixando um desvio-padrão mí-

nimo para modelar a função W, a partir da média dos desvios-padrão das estimativas de intra-variabilidade ou a partir da média dos desvios-padrão das estimativas de *inter-variabilidade* (função B), estas relacionadas à hipótese da defesa. Apenas com nessa última abordagem foi possível obter um sistema de cálculo de LR que apresentava calibração satisfatória.

A avaliação do desempenho do sistema de cálculo de LR foi realizada através de gráficos Tippett [81], conforme Figura 3.7. Neste tipo de gráfico, que apresenta a proporção de casos com "valores de LR maiores que...", são mostradas simultaneamente a curva para os ensaios em que a hipótese H_p é verdadeira e a curva para os casos em que H_d é verdadeira. Idealmente, a curva relacionada a H_p deve estar mais à direita (com valores mais altos de LR) e a curva relacionada a H_d deve estar mais à esquerda (com valores mais baixos de LR).

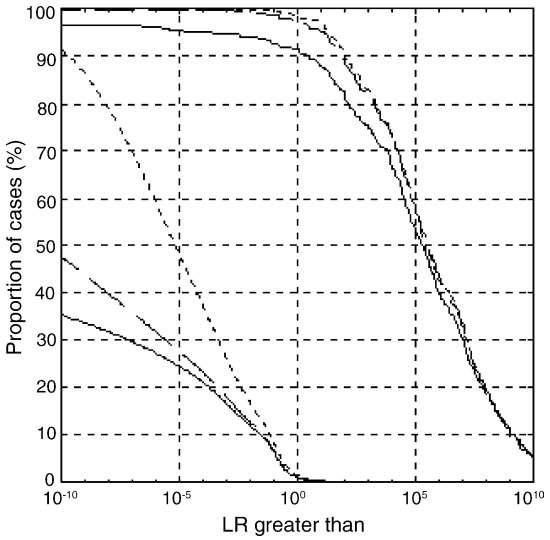


Figura 3.7: Gráficos Tippett dos sistemas de cálculo de LR para reconhecimento facial apresentados em [6]. Reproduzida do original, com permissão.

Na Figura 3.7, as linhas sólidas correspondem ao sistema sem ajuste da intra-variabilidade, resultando em uma alta proporção ($\sim$10%) de casos em que era verdadeira a hipótese da acusação e a LR ficou abaixo de 1, ou seja, teria sido apresentada uma evidência que suporta mais fortemente a hipótese errada. As linhas pontilhadas correspondem à utilização da média de inter-variabilidade como valor mínimo para as estimativas de intra-variabilidade, o que permitiu obter um sistema com melhor calibração.

Abordagem análoga foi apresentada em 2006 por [7], também de forma agnóstica em relação à modalidade biométrica. A abordagem é essencialmente a mesma apresentada no trabalho anterior, porém apresentada com um detalhamento maior, conforme mostrado na Figura 3.8.

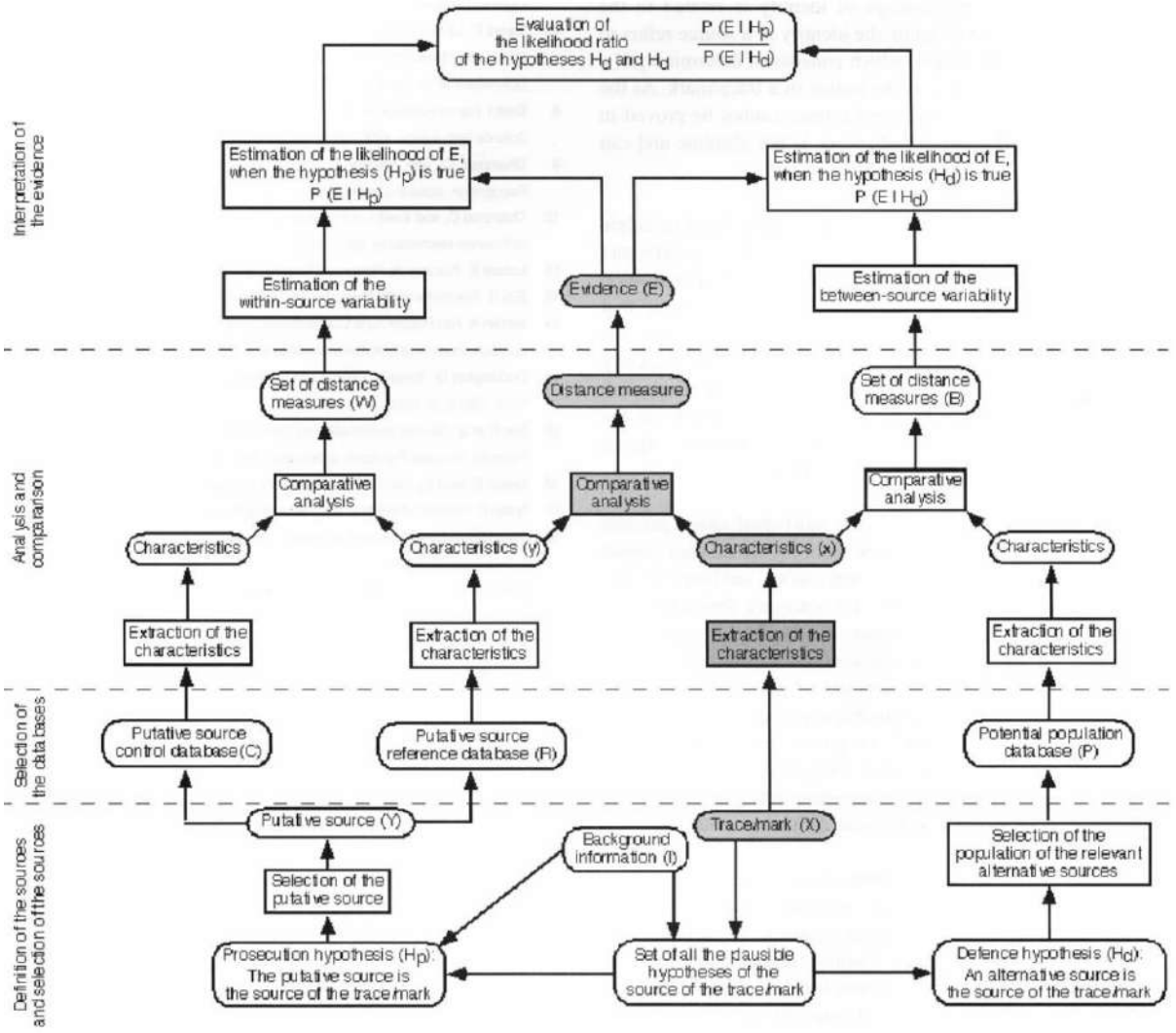


Figura 3.8: Framework proposto em [7] para cálculo de LR a partir de escores de sistemas biométricos. Assim como em [6], H_p é ancorada no suspeito e H_d é ancorada no vestígio. Reproduzida de [7], com permissão.

Em 2014, [82] apresentou uma avaliação do desempenho de verificação e calibração de um sistema de reconhecimento facial com aplicabilidade para o meio forense. Além de ter utilizado a base SCFace, que também será avaliada nesta pesquisa, este trabalho possui dois aspectos que merecem destaque: a utilização de escores que consideram similaridade e tipicidade, através de modelagem ISV (do inglês *inter-session variability*), inspirada na área de reconhecimento de locutor [83], e a utilização da métrica C_{llr} (do inglês *Log-Likelihood-Ratio Cost*) [84] para avaliar o desempenho do sistema.

A C_{llr} é uma das métricas principais a serem utilizadas neste trabalho, pois é apropriada para medir a acurácia e calibração de um sistema de cálculo de LR. A C_{llr} é calculada da seguinte forma:

$$C_{llr} = \frac{1}{2}\left[\frac{1}{N_p}\sum_{i_p} log_2\left(1+\frac{1}{LR_i}\right) + \frac{1}{N_d}\sum_{j_d} log_2(1+LR_j)\right], \tag{3.1}$$

onde N_p e N_d são as quantidades de LRs associadas às hipóteses H_p e H_d, respectivamente; i_p e j_d são os índices das LRs associadas a H_p e a H_d, respectivamente, e LR_i e LR_j são os valores individuais das LRs nos conjuntos associados a H_p e a H_d, respectivamente.

Por ser um custo, ou uma função de perda, o objetivo a ser perseguido é obter sistemas com baixa C_{llr} , sendo útil considerar que um sistema que sempre apresenta LR = 1, ou seja, que nunca oferece suporte maior a uma das hipóteses, possui $C_{\text{llr}} = 1$.

Ao analisar os termos que compõem a equação 3.1, temos que o termo com o primeiro somatório se refere às LR obtidas para os testes em que é verdadeira a hipótese H_p, ou seja, a hipótese da acusação. Assim, este termo penaliza mais fortemente testes que resultaram em valores baixos de LR, quando deveriam ter resultado em valores altos (uma vez que H_p era verdadeira). Alternativamente, o termo com o segundo somatório envolve os testes em que H_d era verdadeira, e neste termo são penalizados os testes que resultaram em LR de valor elevado, quando deveriam ter resultado em valores baixos.

Em 2014 [85] testou combinações variadas de algoritmos/sistemas de reconhecimento facial e de métodos para cálculo de LR a partir de escores. Em que pese os sistemas testados serem relativamente ultrapassados e com desempenho significativamente inferior aos disponíveis atualmente, o trabalho ofereceu contribuições importantes por comparar três métodos distintos de cálculo de LR a partir de escores (KDE - *Kernel Density Estimation*, RL - Regressão Logística e PAV - *Pool Adjacent Violators*) e comparou as abordagens baseadas em hipóteses ancoradas e não-ancoradas.

Dessas comparações, [85] concluiu que KDE e PAV eram mais sensíveis a variações na quantidade de escores utilizados para treinamento, enquanto RL era mais sensível à forma das distribuições de escores de treinamento. Além disso, destacou que KDE possui a propriedade indesejada de que o mapeamento de escores para LR pode não ser monotônico. O autor sugeria ainda que um intervalo de LR fosse reportado, ao invés de um único valor, para considerar efeitos de variabilidade de amostragem.

Em relação à comparação entre as abordagens ancorada e não-ancorada para as hipóteses, foi observado que apenas valores muito altos de LR apresentavam diferenças significativas entre as duas abordagens e que, ao se considerar a apresentação de resultados utilizado escalas de equivalentes verbais para intervalos de LR, as conclusões eram concordantes em 59,2% dos casos. As diferenças encontradas em valores mais altos de LR diminuíam, por outro lado, quando se comparavam sistemas que empregavam a mesma quantidade de escores de treinamento, o que sugere que a escassez de dados para modelar as hipóteses ancoradas é, de fato, um problema prático importante.

Em 2020 [30] desenvolveu e avaliou cinco sistemas para conversão de escore em LR, utilizando dados reais da Autoridade Sueca de Polícia. Neste trabalho, o desempenho dos sistemas de cálculo de LR foi avaliado em termos de C_{llr}, gráficos Tippett e curvas ECE (do inglês *Empirical Cross-Entropy*) [86], além das métricas acurácia, sensitividade e especificidade. Tais critérios de avaliação são essencialmente os mesmos que serão utilizados nesta pesquisa, embora a comparação dos resultados ficará limitada, uma vez que não foram informados ou disponibilizados a base de imagens e o sistema biométrico utilizados.

Em todos os cinco sistemas avaliados, as hipóteses consideradas foram do tipo não-ancorada e os sistemas se diferenciaram em termos da sistemática de conversão de escore para LR. Foram testadas duas abordagens paramétricas para obtenção das PDF que modelam as distribuições de escores sob H_p e sob H_d: distribuições gaussiana normal e gaussiana "skewed". Também foram testados os métodos KDE, Regressão Logística (também avaliados em [85]) e ROCCH (ROC *convex hull*). Este último método é equivalente ao PAV, também examinado por [85]. O sistema de reconhecimento facial utilizado para gerar os escores não foi revelado, sendo apenas descrito pelos autores como "a high performing commercial software".

Os sistemas foram avaliados segundo uma estratégia de validação cruzada 10-fold, na qual, a cada iteração, 90% dos escores são utilizados para treinar o sistema de conversão de escore para LR, e 10% dos escores são utilizados para gerar LRs de teste. Verificou-se que os métodos KDE e gaussiano "skewed" apresentaram maior variabilidade entre as interações da validação cruzada, revelando uma maior sensibilidade aos dados de treinamento para esses dois métodos.

Também foi avaliado o impacto no desempenho em razão do tamanho do conjunto de escores disponíveis para treinamento. Neste aspecto, o método baseado em Regressão Logística foi especialmente sensível à quantidade de escores disponíveis para treinamento, enquanto os demais métodos apresentaram maior estabilidade.

Em 2019, o trabalho [8] apresentou uma revisão da literatura de metodologias para cálculo de LR a partir de escores de reconhecimento facial, além de descrever aplicações de sistemas reconhecimento facial para outras finalidades além da perícia, como investigação e inteligência. Entretanto, o foco foi direcionado para a interpretação da evidência para fins forenses, com destaque para as diferentes possibilidades de modelagem das hipóteses (ancoradas *versus* não-ancoradas).

A Tabela 3.3, adaptada de [8] ilustra as diferentes possibilidades de modelagem das hipóteses, explicita as condicionantes de cada probabilidade a ser obtida para o cálculo da LR e exemplifica o tipo de hipótese considerado.

Conforme mencionado no Capítulo 2, os diferentes tipos de hipótese/ancoragem implicam em formas diferentes de obtenção dos conjuntos de escores utilizados para modelar as

Variabilidade	Ancoragem	Parte relevante na fórmula da LR	Proposição
BSV	no vestígio	$= \frac{num}{\boldsymbol{f(s(S,T)\|T,H_d,I)}}$	H_d: "A fonte do vestígio T não é o suspeito, mas outra pessoa da população de referência"
	no suspeito	$= \frac{num}{\boldsymbol{f(s(S,T)\|S,H_d,I)}}$	H_d: "A fonte do vestígio T não é o suspeito"
	não-ancorada	$= \frac{num}{\boldsymbol{f(s(S,T)\|H_d,I)}}$	H_d: "As imagens são de pessoas diferentes"
WSV	no suspeito	$= \frac{\boldsymbol{f(s(S,T)\|S,H_p,I)}}{denom}$	H_p: "A fonte do vestígio T é o suspeito"
	não-ancorada	$= \frac{\boldsymbol{f(s(S,T)\|H_p,I)}}{denom}$	H_p: "As imagens são de um único indivíduo"

Tabela 3.3: Resumo das diferentes abordagens para modelar as hipóteses H_p (WSV) e H_d (BSV). Nas fórmulas, $\boldsymbol{s}$ representa o escore do caso, $\boldsymbol{S}$ representa o material do suspeito, $\boldsymbol{T}$ representa o vestígio (do inglês *trace*) e $\boldsymbol{f(\cdot)}$ representa a função que modela a densidade de probabilidade relacionada a H_p ou H_d, conforme o caso. Adaptada de [8].

funções densidade de probabilidade relacionadas a H_p e H_d. As Figuras 3.9 e 3.10 ilustram, respectivamente, as diferentes possibilidades de obtenção dos escores para modelagem da inter-variabilidade (BSV), relacionada a H_d, e para modelagem da intra-variabilidade (WSV), relacionada a H_p, segundo Jacquet e Champod [8].

Além disso, em [8] também foi apresentado um *framework* (Figura 3.11) para cálculo de LR a partir de escores, atualizando o que fora apresentado por Meuwly em [7], com a inclusão de modelagens de hipóteses não-ancoradas. Neste *framework*, inicialmente as imagens sob exame são avaliadas quanto à sua adequabilidade para análise por sistemas automáticos, podendo ser decidido que apenas análises manuais por especialistas são possíveis para o caso.

Em seguida, na etapa de especificação do modelo, as hipóteses da acusação e da defesa são determinadas ou identificadas, sendo então delimitada a população de referência. Nesta etapa também é avaliado se as modelagens de intra- e inter-variabilidade serão ancoradas ou não-ancoradas.

A seguir são obtidos os escores para modelagem das hipóteses H_p e H_d, de acordo com a abordagem definida na etapa anterior (ancorada ou não-ancorada), além de se obter o escore (evidência) entre o material do suspeito e o vestígio.

Na última etapa, são estimadas funções densidade de probabilidade a partir dos conjuntos de escores obtidos na etapa anterior, para modelagem de H_p e H_d, e a LR é calculada avaliando-se o valor de cada função de densidade de probabilidade no valor do escore obtido da comparação entre o material do suspeito e o vestígio. No caso de

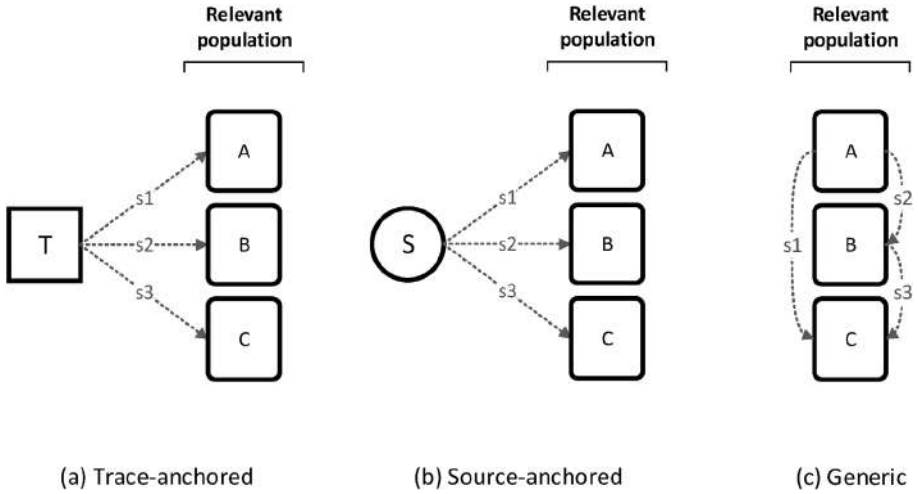


Figura 3.9: Obtenção dos escores para modelagem de H_d. Em (a), os escores são obtidos comparando o vestígio com as amostras da população de referência. Em (b), os escores são obtidos de comparações entre o material do suspeito e as amostras da população de referência. Em (c), os escores são obtidos a partir de comparações apenas entre as amostras da população de referência, não incluindo comparações com materiais do caso. Reproduzida de [8], com permissão.

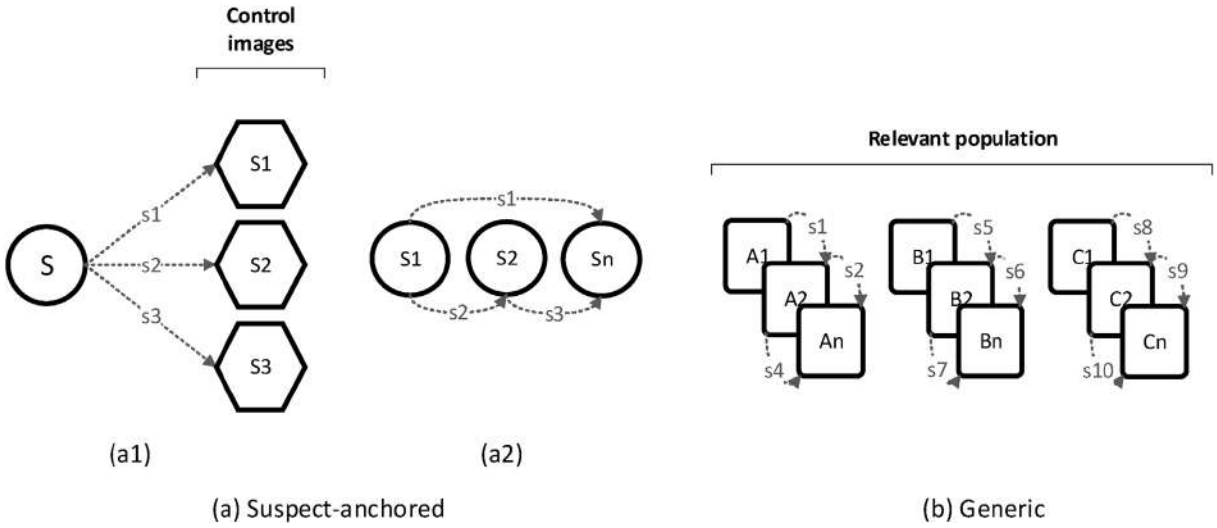


Figura 3.10: Obtenção dos escores para modelagem de H_p. Em (a1), os escores são obtidos comparando-se o material padrão do suspeito com outras imagens do suspeito, obtidas nas mesmas condições do vestígio. Em (a2), os escores são obtidos de comparações entre todos os materiais do suspeito, independentemente de sua condição de aquisição. Em (b), os escores são obtidos a partir de comparações entre as amostras dos indivíduos da população de referência, não incluindo comparações com materiais do caso. Reproduzida de [8], com permissão.

abordagens não-paramétricas, os conjuntos de escores são utilizados para treinamento do sistema de conversão de escore em LR.

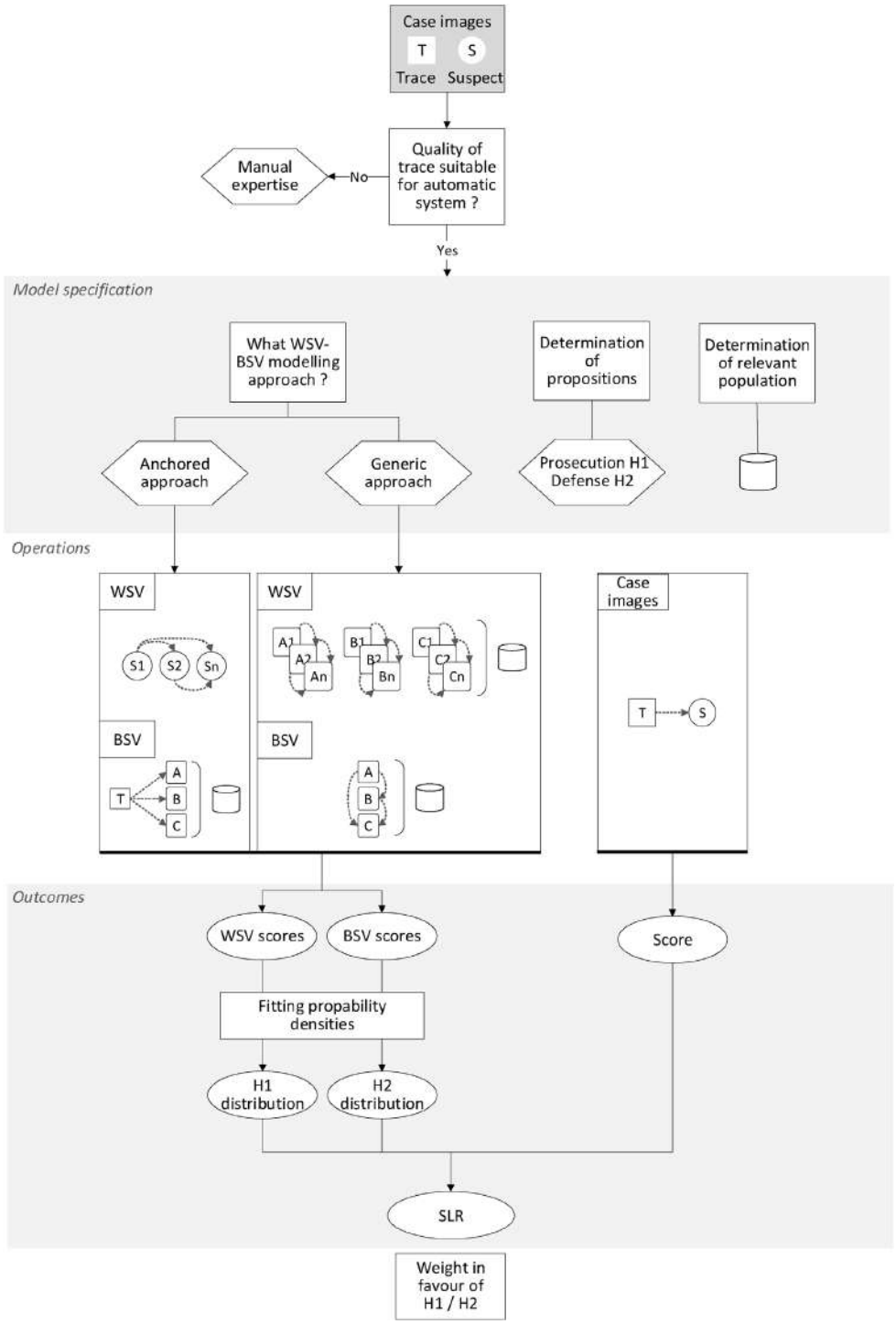


Figura 3.11: Framework proposto por [8] para obtenção de LR a partir de escores. SLR significa *Score-based LR*. Reproduzida de [8], com permissão.

Finalmente, como trabalho mais recente relacionado ao cálculo de LR a partir de sistemas de reconhecimento facial, [87] propôs um método para estimar a qualidade de imagens faciais e incorporar esta informação para o cálculo de LR a partir de escores, com foco em imagens de baixa qualidade.

A estimativa de qualidade da imagem é feita através do que os autores denominaram Escore de Confusão (do inglês *Confusion Score* - CS), que é calculado como o valor médio dos dez maiores escores de similaridade obtidos de comparações entre a imagem de interesse e imagens de outras pessoas que apresentam, também, baixa qualidade.

A partir do cálculo do CS de cada imagem, a LR é calculada de forma similar à descrita no *framework* proposto por Jacquet e Champod, mas selecionando imagens para modelagem de H_p e H_d que possuam CS dentro de um intervalo próximo ao CS da imagem que representa o vestígio.

Como resultados relevantes e de interesse direto desta pesquisa, [87] demonstrou um redução significativa da C_{llr} , apresentada indiretamente através de curvas ECE [86], ao comparar um sistema que considera o CS para selecionar imagens para modelagem de H_p e H_d, relativamente a outro sistema que não utiliza o valor CS para essa modelagem.

3.4.1 Limitando o valor da LR

Em situações em que o cálculo da LR é realizado a partir de modelos obtidos de poucas de amostras, existe a possibilidade de que as LR obtidas estejam muito subestimadas ou muito superestimadas [10, 12].

Duas alternativas foram propostas na literatura para aliviar este problema: ELUB [10] e regressão logística regularizada [12].

Em [10] foi proposto um método para impor limites superiores e inferiores às LR obtidas a partir de escores. Não se trata de um método de cálculo de LR *per se*, mas sim de um procedimento adicional que limita as magnitudes das LR obtidas e que pode, por tanto, ser utilizado em conjunto com qualquer dos métodos de cálculo de LR.

A justificativa para a limitação dos valores de LR parte da observação de que valores muito altos ou muito baixos de LR são obtidos em regiões em que uma das distribuições de escores é avaliada em sua região de cauda. Naturalmente, na região da cauda da distribuição é esperado que poucos dados estejam disponíveis e, além disso, pequenas variações dos parâmetros utilizados para estimar a função densidade de probabilidade ou a própria escolha da função para modelar as distribuição podem acarretar em alterações importantes nos valores dessa função na região de cauda, impactando no valor da LR obtida. A Figura 3.12 exemplifica a situação.

Em situações em que poucos dados estão disponíveis para modelar as distribuições de escores, é esperado que os parâmetros obtidos, como média e desvio-padrão, apresentem efeitos de variabilidade de amostragem importantes, sendo então justificada a limitação dos valores de LR.

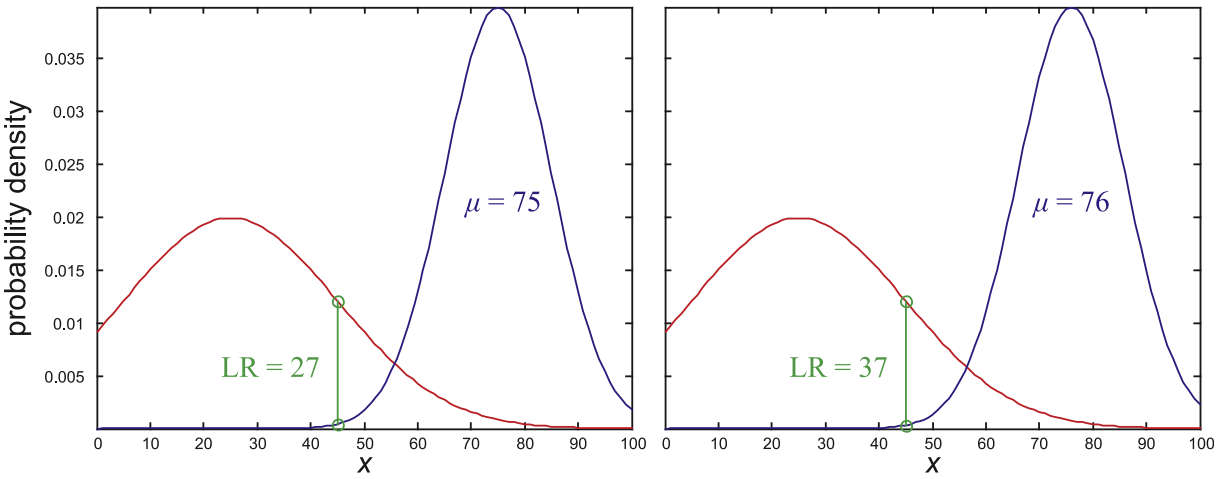


Figura 3.12: Variabilidade da LR devido a pequenas variações nos parâmetros utilizados para modelar as distribuições de escores. Reproduzida de [9], com permissão.

O método ELUB é basead no conceito de *Normalized Bayes Error-rate* - NBE [88] que, por sua vez, é fundamentado em teoria de decisão aplicada a decisões binárias. Em processos que envolvem decisões deste tipo, NBE considera quatro possíveis resultados:

- condenar um inocente (ci)
- condenar um culpado (cc)
- inocentar um inocente (ii)
- inocentar um culpado (ic)

A perspectiva de decisões binárias no contexto desta pesquisa se relaciona com as decisões a serem tomadas pela instância julgadora a partir da LR apresentada no exame de comparação facial. De forma simplificada, desconsideramos a existência de outras evidências e supomos que as hipóteses consideradas para o cálculo da LR se relacionam diretamente com as hipóteses de culpa e inocência a serem avaliadas pelo juízo.

Em teoria de decisão, uma função de utilidade é definida a partir da atribuição de utilidades (ou custos, em interpretação inversa) associadas a cada um dos quatro resultados descritos - U_{ci}, U_{cc}, U_{ii} e U_{ic}, respectivamente. Considerando decisões racionais, o objetivo em um processo de tomada de decisão seria sempre o de maximizar a função de utilidade que envolve esses quatro possíveis resultados. Assim, pode-se definir a utilidade esperada (*Expected Utility* - EU) como:

$$EU = max\{U_{cc} \times P(H_p) + U_{ci} \times (1 - P(H_p)); U_{ic} \times P(H_p) + U_{ii} \times (1 - P(H_p))\}, \quad (3.2)$$

onde $P(H_p)$ é a probabilidade da hipótese da acusação ser verdadeira.[1]

Assim, a decisão pela hipótese da acusação, ou seja, por condenar, ocorreria na situação em que $U_{cc} \times P(H_p) + U_{ci} \times (1 - P(H_p)) > U_{ic} \times P(H_p) + U_{ii} \times (1 - P(H_p))$, enquanto que a decisão de inocentar ocorreria quando $U_{cc} \times P(H_p) + U_{ci} \times (1 - P(H_p)) < U_{ic} \times P(H_p) + U_{ii} \times (1 - P(H_p))$.

É possível rearranjar os termos das equações acima e obter:

- julgar como culpado se $\frac{P(H_p)}{1-P(H_p)} > \frac{U_{ii}-U_{ci}}{U_{cc}-U_{ic}}$
- julgar como inocente se $\frac{P(H_p)}{1-P(H_p)} < \frac{U_{ii}-U_{ci}}{U_{cc}-U_{ic}}$

Nota-se que o termo $\frac{P(H_p)}{1-P(H_p)}$ é a razão de probabilidades a posteriori do caso. Como visto na 2.2, esse termo é o produto da LR pela razão de probabilidades a priori. Assim, vemos que o critério de decisão descrito leva em conta a razão de probabilidades a priori, o valor da LR e as utilidades descritas anteriormente nessa seção.

Embora intuitivo, esse resultado permite estabelecer um valor limiar para a LR, acima do qual a condenação seria pela condenação e abaixo do qual a decisão seria pela inocência. Esse valor, obtido através de manipulações algébricas simples das equações acima, é dado por:

$$LR_{th} = \frac{U_{ii} - U_{ci}}{U_{cc} - U_{ic}} \times \frac{H_d}{H_p} \tag{3.3}$$

Como estamos interessados em avaliar se um determinado sistema de cálculo de LR deve ou não ser utilizado, calculamos a utilidade esperada do sistema. Para isso escolhemos valores para as utilidades de cada resultado como $U_{ii} = U_{cc} = 0$, $U_{ic} = -1$ e $U_{ci} = -\frac{U_{ii}-U_{ci}}{U_{cc}-U_{ic}}$. Segundo [88], estas escolhas resultariam em decisões ótimas. Assim, a EU de um sistema de cálculo de LR pode ser expressa por:

$$EU(sistema\,LR) = -P(H_p) \times P(LR \leq LR_{th}|H_p) - \frac{U_{ii} - U_{ci}}{U_{cc} - U_{ic}} \times P(H_d) \times P(LR > LR_{th}|H_d) \tag{3.4}$$

De forma análoga, um sistema neutro, ou seja, que sempre apresenta como resultado LR = 1, teria como EU:

$$EU(neutro) = -P(H_p) \times P(1 \leq LR_{th}|H_p) - \frac{U_{ii} - U_{ci}}{U_{cc} - U_{ic}} \times P(H_d) \times P(1 > LR_{th}|H_d) \tag{3.5}$$

Assim, é possível avaliar se um determinado sistema é melhor do que um sistema neutro normalizando a utilidade esperada do sistema em relação à utilidade esperada de um sistema neutro, dividindo a equação 3.4 pela equação 3.5. Para facilitar a interpretação dos resultados, essa normalização é invertida e, após manipulações, chega-se a:

[1] Deve-se notar que esta probabilidade é condicionada à evidência, mas a notação foi simplificada nesta seção por questões de diagramação.

$$\frac{EU(neutro)}{EU(sistema\,LR)} = \frac{P(1 \leq LR_{th}|H_p) + LR_{th} \times P(1 > LR_{th}|H_d)}{P(LR \leq LR_{th}|H_p) + LR_{th} \times P(LR > LR_{th}|H_d)} \tag{3.6}$$

Uma vez que, na equação 3.6, o termo LR_{th} é desconhecido do perito, ele é tomado como variável independente, sendo obtido um gráfico que representa o NBE em função de LR_{th}, em escala logarítmica. A Figura 3.13 mostra um exemplo de um gráfico NBE para um sistema hipotético de cálculo de LR, onde é possível verificar que em algumas regiões de LT_{th} o sistema apresenta utilidade inferior ao sistema neutro, ou seja, a utilização do sistema é prejudicial à tomada de decisões do julgador.

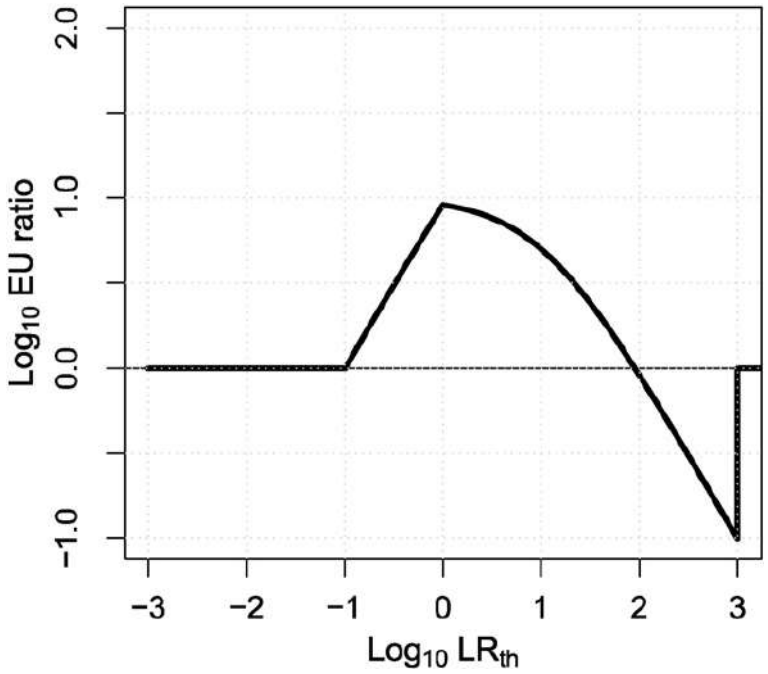


Figura 3.13: Gráfico NBE de um sistema hipotético. Reproduzida de [10], com permissão.

Assim, segundo o método ELUB, são definidos os valores mínimo e máximo de LR que poderiam ser obtidos do sistema, de modo a que, em nenhum caso, as LR obtidas do sistema apresentem utilidade inferior à de um sistema neutro. No exemplo da Figura 3.13 esses valores seriam, aproximadamente -1 e +2 (em escala log_{10}). LRs obtidas desse sistema com valores menores ou maiores do que esses seriam substituídas pelos valores limites correspondentes.

Embora teoricamente bem fundamentado, a utilização do ELUB na prática possui algumas dificuldades. A primeira é o estabelecimento de limites rígidos para os valores mínimos e máximos da LR, fazendo com que, por exemplo, um escore de valor muito elevado resulte numa LR de mesmo valor da obtida de um escore bastante inferior, mas que ainda esteja acima do limite superior estabelecido pelo ELUB. Além disso, geralmente a quantidade de escores relacionados a H_p é significativamente inferior à de escores re-

lacionados a H_d. Ocorre que a quantidade de escores relacionados a cada hipótese tem influência significativa no cálculo dos valores limite da LR utilizando ELUB. Assim, é comum que o limite superior das LR seja desproporcionalmente maior do que o limite inferior. Esta situação é especialmente crítica em cenários em que poucos dados estão disponíveis e, para a obtenção de modelos estatísticos mais robustos, técnicas de reamostragem são empregadas. Em razão destas dificuldades práticas neste trabalho optou-se por não avaliar o método ELUB. Esta avaliação é deixada como trabalho futuro, em que as dificuldades práticas descritas neste parágrafo serão também examinadas.

A regressão logística regularizada será discutida na Seção 4.3.3.

Capítulo 4

Materiais e Métodos

Este capítulo descreve os métodos propostos e avaliados nesta pesquisa, as bases de imagens faciais utilizadas e os critérios de validação.

Inicialmente, cabe esclarecer que foram adotadas as seguintes restrições e extensões em relação ao *framework* proposto por [8], descrito na Figura 3.11:

- Apenas hipóteses não-ancoradas (ou genéricas) são consideradas. A opção por esta restrição se justifica, principalmente, pela dificuldade prática de obter amostras do suspeito em condições equivalentes às do vestígio, o que é necessário para modelar adequadamente hipóteses ancoradas. Além disso, os resultados relatados por [6] e [85] indicam que a baixa quantidade de dados geralmente disponíveis para modelar hipóteses da acusação do tipo ancorada pode ser um problema maior do que o caráter genérico das hipóteses não-ancoradas.

- Inclusão possibilidade de agregação de *embeddings* antes do cálculo dos escores, para tratamento de vídeos ou de casos em que estejam disponíveis múltiplas imagens do suspeito.

- Avaliação de método com limitação dos valores de LR, através da regularização da regressão logística.

4.1 Sistemas de Reconhecimento Facial

Foram utilizados sistemas biométricos constituídos por algoritmos e modelos da literatura com implementações em código aberto.

O primeiro sistema é baseado no modelo FaceNet, publicado originalmente em 2015 [24] e cuja implementação utilizada nesta pesquisa foi a disponibilizada pela biblioteca

DeepFace[1] / LightFace [89]. Neste sistema o estágio de detecção facial é baseado no método *Multi-Task Cascaded Convolutional Neural Networks* (MTCNN) [90].

O segundo sistema é baseado no modelo ArcFace, publicado originalmente em 2019 [54] e implementado pela biblioteca InsightFace[2]. Nesta biblioteca o estágio de detecção facial é baseado no método *Sample and Computation Redistribution for Efficient Face Detection* (SCRFD) [91].

4.1.1 FaceNet

O modelo baseado em FaceNet foi treinado na base VGGFace2 e incluído na biblioteca DeepFace / LightFace. Foi reportado pelo autor da biblioteca que este modelo obteve acurácia de 99,2% na base LFW, segundo o protocolo *Unrestricted, Labeled Outside Data*. O modelo é baseado na arquitetura InceptionV1 e produz *embeddings* com 128 dimensões.

4.1.2 ArcFace

O modelo baseado em ArcFace foi treinado no âmbito desta pesquisa, utilizando para treinamento a base MS1M-Arcface [54], disponibilizado pelos autores da biblioteca InsightFace e contendo aproximadamente 5,8 milhões de imagens faciais de 87 mil indivíduos distintos.

Este modelo obteve acurácia de 99,83% na base LFW, segundo o protocolo *Unrestricted, Labeled Outside Data*, resultado compatível com o estado da arte. O modelo é baseado na arquitetura ResNet e produz *embeddings* com 512 dimensões.

Ressalta-se que, diferentemente dos modelos pré-treinados disponibilizados pelos autores da biblioteca InsightFace, o modelo treinado e utilizado nesta pesquisa não possui restrições de licenciamento para uso profissional. Assim, sua utilização é permitida a profissionais de perícia para a realização de exames periciais segundo as técnicas descritas e validadas nesta pesquisa.

4.2 Bases de imagens faciais utilizadas

4.2.1 FEI

Conforme visto na Seção 3.2, esta base é composta por um total de 2.800 imagens de 199 indivíduos. Foi selecionado um subconjunto dessa base, composto por imagens em pose aproximadamente frontal, de forma a obter um conjunto que se aproximasse das condições

[1] `https://github.com/serengil/deepface`
[2] `https://github.com/deepinsight/insightface`

de exames periciais em que são avaliadas imagens de bancos de identificação. A Figura 4.1 ilustra o subconjunto das imagens que foram selecionadas de cada indivíduo.

Figura 4.1: Imagens selecionadas (em vermelho) da base FEI.

Além disso, as cinco imagens de cada pessoa foram processadas para simular as condições típicas de imagens usualmente presentes nas bases de identificação, adotando-se a seguinte sistemática:

- duas imagens foram mantidas sem alterações;
- uma imagem foi reduzida para 1/3 da resolução original e em seguida redimensionada de volta para a resolução original, utilizando interpolação por vizinho mais próximo;
- uma imagem teve ruído gaussiano adicionado, com os seguintes parâmetros: média igual a zero, desvio-padrão igual a 0,3 e intensidade relativa igual a 0,2; e
- uma imagem foi filtrada com aplicação de filtro *unsharp mask*, com os parâmetros: quantidade igual a 3, raio igual a 5 e limiar igual a 0.

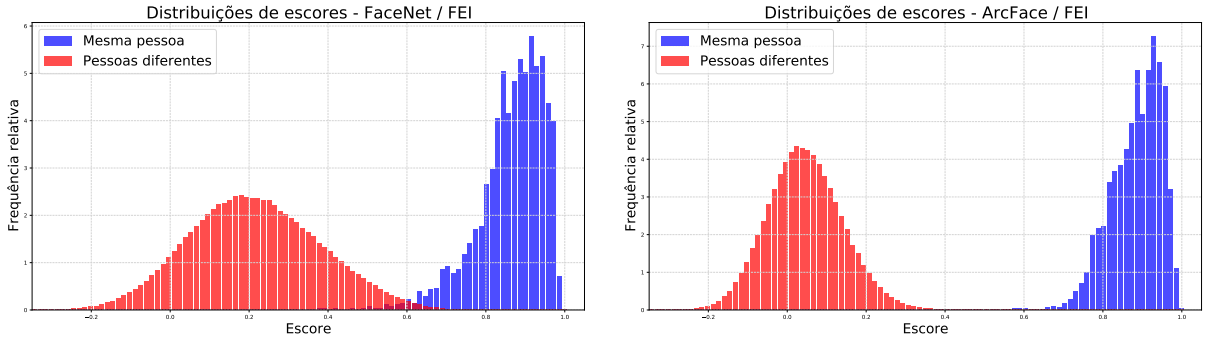
Os parâmetros indicados acima foram obtidos por ensaios e inspeção visual dos resultados, levando-se em conta a experiência do autor com o tipo de imagem usualmente recebido para perícias.

As imagens serão comparadas duas a duas, resultando, para cada sistema em 2.025 escores obtidos de comparações de imagens de um mesmo indivíduo e 497.475 escores obtidos de comparações de imagens de indivíduos diferentes.

Os histogramas desses dois conjuntos de escores são mostrados nas Figuras 4.2a, para o FaceNet, e 4.2b, para o ArcFace.

4.2.2 SCFace

A base SCFace contém imagens de 130 pessoas, distribuídas da seguinte forma:



(a) (b)

Figura 4.2: Distribuições de escores na base FEI obtidas com o FaceNet (à esquerda) e com o ArcFace (à direita).

- 9 imagens de excelente qualidade, capturadas por câmeras fotográficas, com variados graus de rotação, desde perfil lateral esquerdo, até perfil lateral direito, incluindo uma fotografia frontal;
- 15 imagens capturadas por câmeras de videomonitoramento, dispostas a 2,25 m de altura e capturadas a três distâncias entre o indivíduo e as câmeras, horizontalmente: 1,0 m, 2,6 m e 4,2 m; e
- outras imagens em modo infra-vermelho, que não são de interesse desta pesquisa.

Conforme comentado na Seção 3.2, tanto esta base quando a Quis-Campi são úteis para avaliar a viabilidade de utilização de sistemas de reconhecimento facial para exames periciais em imagens oriundas de câmeras de videomonitoramento, que apresentam desafios significativos para o reconhecimento de indivíduos.

A título ilustrativo, a Figura 4.3 exibe as distribuições de escores obtidas na base SCFace, utilizando o modelo baseado em ArcFace para comparar as imagens de fotografia frontal com imagens das câmeras de videomonitoramento na distância 1, ou seja, a 4,2 m.

É possível verificar na Figura 4.3 que há muito mais sobreposição entre as distribuições de escores do que no caso da Figura 4.2b, obtida da base FEI, que apresenta imagens com qualidade superior às da base SCFace.

4.2.3 Quis-Campi

Assim como a base SCFace, a base Quis-Campi também permite avaliar o emprego de sistemas biométricos para cálculo de LR em imagens de baixa resolução. O principal diferencial desta base em relação à SCFace é que as imagens das câmeras de videomonitoramento foram capturadas em ambiente externo, com iluminação não controlada e com os indivíduos em deslocamento. Há também um número maior de indivíduos registrados

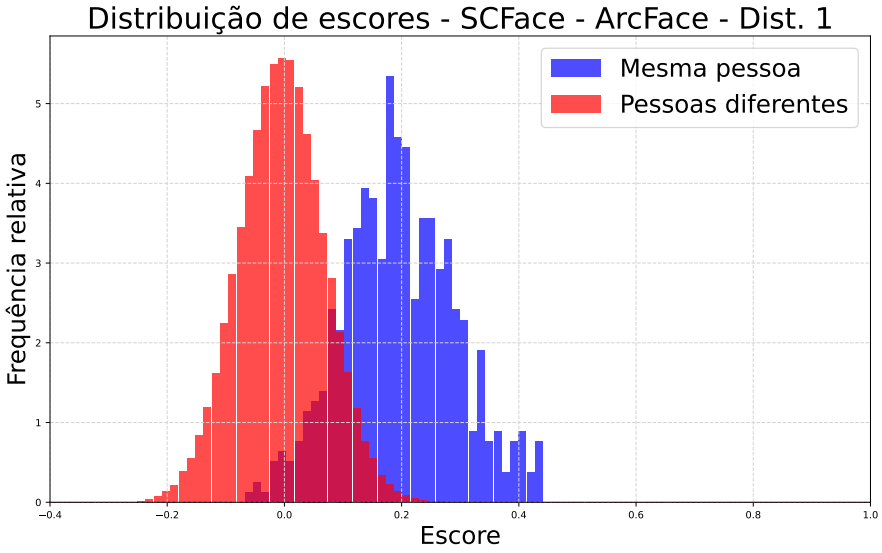


Figura 4.3: Distribuições de escores na base SCFace, obtidas com ArcFace para imagens com pior resolução.

(320), sendo este outro grande atrativo desta base, dada a dificuldade em se obter bases com esse tipo de imagem.

4.2.4 Adience

A base Adience foi apresentada em [68] e é composta por 26.580 imagens de2.284 indivíduos distintos. As imagens foram coletadas de repositórios de imagens na Internet e têm origem em telefones celulares ou outros dispositivos móveis de captura de imagem. As imagens apresentam variações significativas de pose, ilumincação, expressão facial e outros fatores relacionados à qualidade das imagens. Considerando que o número de imagens por identidade nesta base é altamente desbalanceado, um subconjunto de imagens foi escolhido para os experimentos, incluindo apenas imagens de identidades em que estão disponíveis pelo menos 11 imagens. Esta seleção resultou em um conjunto de 14.143 imagens de 373 identidades. Esta base será útil para avaliar o cenário de imagens de redes sociais e foi utilizada nesta pesquisa apenas nos experimentos relacionados a agregação de *embeddings* descritos na Seção 4.4.

4.2.5 BFW

A base BFW foi introduzida em 2020 e é composta por 20.000 imagens de 800 indivíduos, divididas em oito subgrupos, correspondendo às combinações de grupos étnicos (Asiáticos, Brancos, Indicanos e Pretos) e gêneros (Feminino e Masculino). As imagens são do tipo

(a) Adience (b) BFW

Figura 4.4: Exemplos de referências selecionadas para as bases Adience e BFW. Para cada identidade, a face acima e à esquerda (em verde) foi selecionada como referência, enquanto as demais são utilizadas como imagens questionadas.

in the wild, com variações de pose, iluminação, expressão e resolução, além das imagens apresentarem graus variados de oclusão.

Conforme adiantado na Seção 3.2, esta base é útil para avaliar métodos de cálculo de LR a partir de escores de sistemas de reconhecimento facial aplicáveis a casos cujas imagens tenham origem em redes sociais ou outra fonte de imagens *in the wild*.

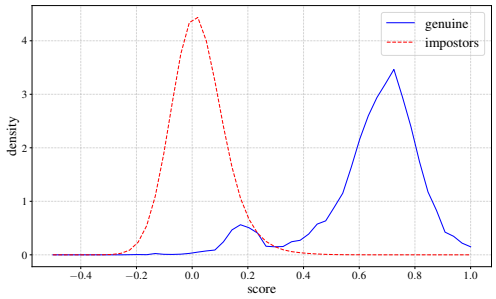
4.2.6 Definição das imagens de referência nas bases Adience e BFW

Uma vez que as bases Adience e BFW não possuem imagens de referência de cada identidade, que poderiam ser utilizadas para simular o material padrão em cenários forenses, foi necessário escolher uma imagem de cada identidade dessas bases para ser utilizada como referência, enquanto as demais são utilizadas para simular as imagens questionadas.

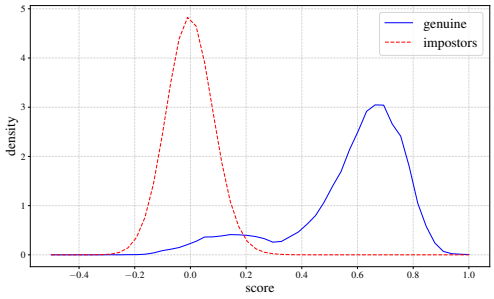
Considerando que nos cenários forenses as imagens de referência são de melhor qualidade, capturadas em condições controladas de pose, iluminação e expressão, foi adotado como parâmetro a escolha da imagem de melhor qualidade de cada identidade como a imagem de referência. Para isso, foi selecionada como referência a imagem cada identidade com melhor ranking combinado de escores de qualidade Ser-Fiq [92] e *Confusion Score* [93]. A Figura 4.4 ilustra a seleção de referências para quatro identidades dessas bases, exemplificando que este critério induziu a seleção de referências de boa qualidade.

4.2.7 Erros de identidade nas bases Adience e BFW

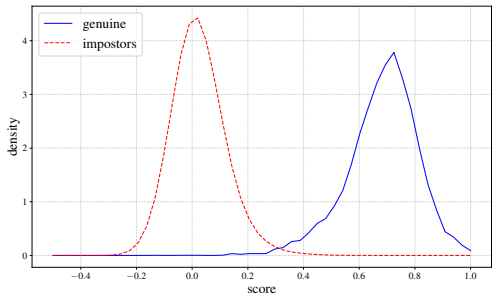
Durante os experimentos iniciais com as bases Adience e BFW, verificou-se que as distribuições de escores de mesmas pessoas (Figuras 4.5 (a) e (b)) apresentam comportamento bimodal, o que levantou suspeitas de que poderiam haver erros nos rótulos de identidade de algumas imagens



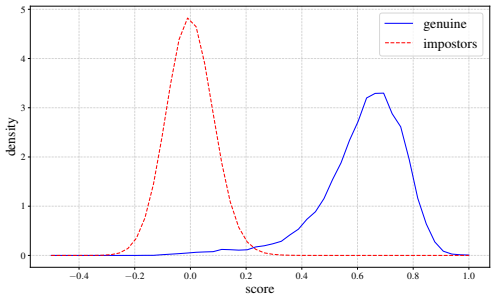
(a) Adience



(b) BFW



(c) Adience clean

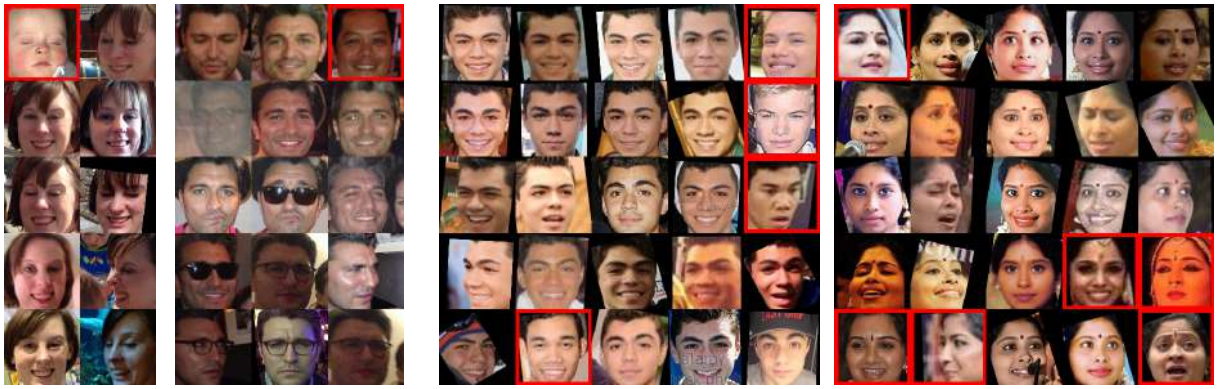


(d) BFW clean

Figura 4.5: Comportamento bimodal das distribuições de escores SS para as bases Adience (a) e BFW (b), sugerindo a existência de erros em rótulos de identidade. Após a limpeza das bases, as distribuições de escores SS não apresentam mais o comportamento bimodal (c, d).

De fato, após inspeção visual das imagens associadas com maior frequência a escores de mesmas pessoas valor mais baixo, foram identificados diversos casos de erros em rótulos de identidade - imagens de identidades distintas associadas a um mesmo indivíduo. A Figura 4.6 mostra alguns exemplos desses erros.

Foi aplicada técnica para correção desse tipo de erro, proposto por [94], resultando em novas versões dessas bases, denominadas *Adience clean*, contendo 13.160 imagens de 355 indivíduos, e *BFW clean*, contendo 19.131 imagens de 800 identidades. No caso da base Adience, foram identificadas ainda 841 arquivos duplicados na base, com o mesmo resumo criptográfico. Estes arquivos também foram removidos para a obtenção da base Adience clean.

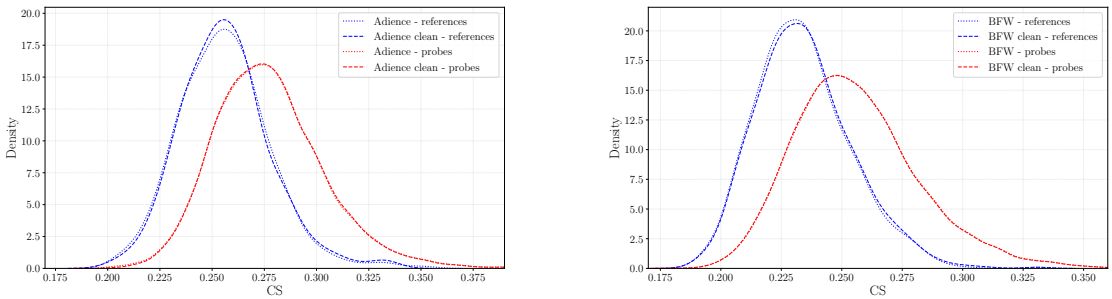


(a) Adience (b) BFW

Figura 4.6: Exemplos de erros nos rótulos de identidade (em vermelho) nas bases Adience e BFW.

Para avaliar a eficácia do processo de limpeza, foram observadas as diferenças entre as distribuições dos escores SS e DS antes e depois de limpar bases. As distribuições de escores SS de ambos os conjuntos de dados limpos apresentam uma distribuição uni-modal mais típica, Figuras 4.5 (c) e (d), indicando que o processo de limpeza obteve sucesso em determinar as imagens com rótulos incorretos.

Para avaliar se o procedimento de limpeza alterou a dificuldade do reconhecimento facial dessas bases, avaliamos as diferenças nas distribuições dos escores de qualidade *confusion scores* das imagens de referência e daquelas que foram utilizadas como questionadas antes e depois da limpeza. Conforme ilustrado na Figura 4.7, as distribuições dos scores de qualidade antes e depois do processo de limpeza são muito semelhantes, sugerindo que o procedimento de limpeza não alterou a dificuldade intrínseca dessas bases.



(a) Adience (b) BFW

Figura 4.7: Distribuição de escores de qualidade *consufion scores* para as imagens de referência e questionadas das bases Adience e BFW, antes e após o processo de limpeza.

4.3 Métodos para cálculo de LR a partir de escores

Conforme visto na Seção 3.4, os *frameworks* propostos para o cálculo de LR a partir de escores consideram que a evidência a ser avaliada no paradigma da LR é o escore obtido entre o vestígio e o material do suspeito. Ou seja, no caso de reconhecimento facial, é o escore obtido da comparação entre duas imagens.

Nos dois sistemas biométricos utilizados nesta pesquisa (FaceNet e ArcFace) o escore é calculado como a similaridade cosseno entre as *embeddings* obtidas de cada face sendo comparada. A similaridade cosseno é calculada por:

$$s(\boldsymbol{t}, \boldsymbol{s}) = \frac{\boldsymbol{t} \cdot \boldsymbol{s}}{||\boldsymbol{t}||\,||\boldsymbol{s}||} \tag{4.1}$$

Na Equação 4.1, s representa o escore, $\boldsymbol{t}$ representa as *embeddings* da face no material questionado (*trace*), $\boldsymbol{s}$ representa as embeddings do material do suspeito, $\cdot$ representa o produto interno e $||\,||$ representa a norma do vetor. Essa forma de calcular o escore possui uma interpretação geométrica, representando o cosseno do ângulo entre os vetores $\boldsymbol{t}$ e $\boldsymbol{s}$.

Para calcular LR a partir do escore, é preciso estimar o quanto o escore é plausível supondo verdadeira a hipótese da acusação ($p(s|H_p)$) e também o quanto o mesmo escore é plausível supondo verdadeira a hipótese da defesa ($p(s|H_d)$).[3] Uma vez obtidas essas estimativas, a LR é obtida pela razão entre o primeiro valor e o segundo, conforme exposto na Equação 2.1.

O cálculo do numerador e do denominador da Equação 2.1 pode ser feito de diversas formas, que podem ser agrupadas em métodos paramétricos, em que são utilizadas funções densidade de probabilidade com parâmetros estimados a partir das distribuições de escores correspondentes a H_p e a H_d, e métodos não paramétricos, em que a estimativa das funções densidade de probabilidade não depende da estimação de parâmetros dos conjuntos de escores. No caso desta pesquisa, a estimativa por densidade de *kernel* (KDE) é um método que será utilizado e se enquadra nesta categoria.

Há ainda métodos não paramétricos que não dependem do cálculo explícito do numerador e do denominador da Equação 2.1. Nesta pesquisa, a Regressão Logística é um método que será utilizado e possui esta característica. O cálculo de LR a partir do escore é feito com a aplicação de uma função logística cujos parâmetros são aprendidos a partir dos conjuntos de escores relacionados a H_p e a H_d.

[3] O termo referente à informação de contexto foi omitido apenas para simplificar a notação, mas o leitor deve sempre considerar que a informação de contexto condiciona todas as probabilidades calculadas neste trabalho.

4.3.1 Métodos paramétricos

Como visto na seção anterior, métodos paramétricos são baseados em funções densidade de probabilidade conhecidas e cujos parâmetros, como média e desvio-padrão, são obtidos dos conjuntos de escores relacionados a H_p e a H_d. A escolha da função para modelar cada hipótese frequentemente é feita através da inspeção visual dos histogramas de escores de cada conjunto e a correção desta escolha pode ser avaliada indiretamente, comparando-se o desempenho obtido de sistemas baseados em modelagens feitas por diferentes tipos de função. Alternativamente, testes estatísticos podem ser utilizados para orientar a escolha da distribuição que melhor modela os dados.

Gaussiana Normal

Distribuições do tipo Gaussiana Normal são relativamente comuns em conjuntos de escores relacionados a H_d, ou seja, aos escores obtidos de imagens de pessoas diferentes. Como exemplo, as distribuições de escores relacionados a H_d mostradas nas figuras 4.2b, 4.3 e 4.5 parecem obedecer a uma distribuição deste tipo. Por outro lado, as distribuições de escores relacionados a H_p frequentemente não são bem modeladas por distribuições do tipo Gaussiana Normal. O impacto da violação dessa condição de normalidade será evidenciado para algumas combinações de base e modelo de reconhecimento, conforme detalhado no Capítulo 5.

4.3.2 Métodos não-paramétricos

Estimativa por Densidade de Kernel

A Estimativa por Densidade de Kernel (KDE) pode ser compreendida como uma técnica para estimar uma função densidade de probabilidade através da suavização do histograma. Embora tenha a vantagem de não assumir *a priori* nenhuma forma específica para a distribuição que se pretende modelar, é um processo sensível a um dos parâmetros utilizados para a suavização do histograma: a largura de banda h.

Considerando um conjunto de amostras independentes e identicamente distribuídas $(x_1, x_2, ..., x_n)$ obtidas de uma distribuição f que se deseja modelar, a estimativa $\hat{f}$ obtida por esta técnica é calculada por:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{4.2}$$

em que K é a função *kernel* utilizada.

Valores muito baixos para h tendem a fazer com que a função estimada fique sobre-ajustada (*overfitted*) às amostras, enquanto valores muito altos de h tendem a fazer com que a função estimada fique sub-ajustada (*underfitted*) aos dados. Alguns autores sugerem algumas regras práticas para escolha de valores adequados de h, sendo a mais conhecida a chamada regra de *Silverman* [95], dada pela seguinte equação e sugerida para modelar distribuições unimodais aproximadamente normais:

$$h = 0,9\, min\left(\hat{\sigma}, \frac{IQR}{1,34}\right) n^{-\frac{1}{5}}, \tag{4.3}$$

onde $\hat{\sigma}$ é o desvio padrão das amostras e IQR é o intervalo interquartil.

De forma geral, a utilização de KDE para modelar distribuições do tipo *heavy-tailed* é considerada difícil [96], pois valores adequados de h para modelar a região com maior densidade geralmente produzem *overfitting* na região da cauda da distribuição, enquanto valores adequados de h para modelar a região da cauda tendem a produzir *underfitting* na região com maior densidade.

Regressão Logística

Regressão logística é uma técnica comumente utilizada para classificação de problemas binários e que pode ser treinada de modo supervisionado. É baseada na utilização da função logística, definida por:

$$f(x) = \frac{1}{1 + e^{-x}}, \tag{4.4}$$

em que $f(x)$ representa a probabilidade *a posteriori* da classe positiva e a probabilidade da classe negativa pode ser obtida por $1 - f(x)$. Ao se considerar probabilidades *a priori* iguais para H_p e para H_d, a razão das probabilidades obtida pela regressão logística pode ser interpretada como LR. Essa consideração de probabilidades *a priori* iguais é feita apenas para o treinamento do modelo de regressão logística, não devendo ser confundida com as probabilidades *a priori* para cada caso em que o sistema venha a ser utilizado.

Para fins de cálculo de LR utilizando este modelo, são utilizados como dados de treinamento dois conjuntos de escores, um deles correspondente à hipótese da acusação, ou seja, as imagens são de uma mesma origem, e o outro correspondente à hipótese da defesa, ou seja, as imagens correspondem a pessoas distintas. O objetivo do treinamento é aprender um mapeamento de escore para probabilidade da hipótese de mesma origem, baseado na função logística. O processo é exemplificado na Figura 4.8.

A utilização de Regressão Logística para cálculo de LR é frequente na área de comparação de locutor [97, 98, 99, 100] e apresenta algumas vantagens importantes em relação aos métodos paramétricos e ao KDE.

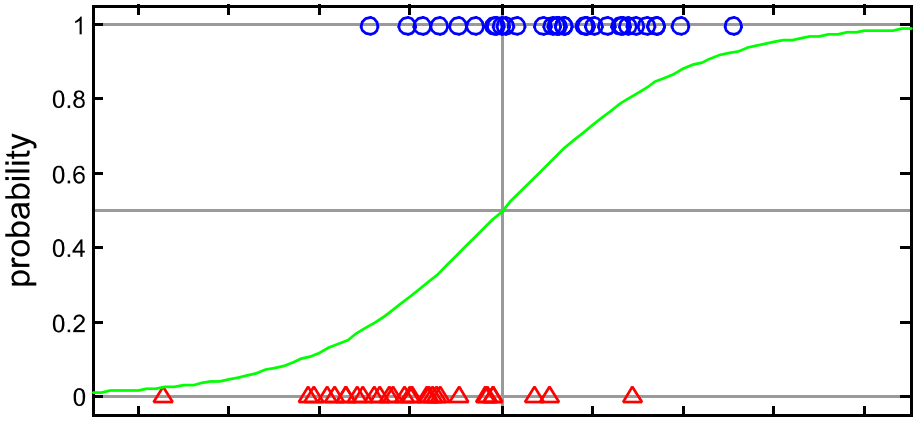


Figura 4.8: Treinamento do modelo de regressão logística. Aos escores de treinamento correspondentes a H_p (em azul) são atribuídos a probabilidade 1 e aos escores de treinamento correspondentes a H_d (em vermelho) são atribuídos a probabilidade 0. A curva correspondente ao modelo treinado, em verde, é obtida de forma a obter o melhor ajuste aos dados de treinamento para uma função logística. Extraída de [11], com permissão

Comparativamente aos métodos paramétricos, nenhuma assunção precisa ser feita sobre as funções que modelam as distribuições de escores. Em relação ao KDE, a Regressão Logística é menos sensível a *outliers* [11].

Por outro lado, a utilização de Regressão Logística apresenta dificuldades quando os conjuntos de escores de treinamento apresentam pouca ou nenhuma sobreposição (por exemplo, na Figura 4.2b). Nesses casos, estratégias de regularização como as discutidas em [101] podem ser apresentadas para aliviar o problema.

4.3.3 Regressão Logística Regularizada

Conforme mencionado no final da sub-seção 4.3.2, o treinamento de modelos de regressão logística em conjuntos de dados que apresentam pouca ou nenhuma sobreposição entre as classes apresenta dificuldades práticas. A falta de sobreposição dos conjuntos de treinamento é ainda mais possível de ocorrer quando se tem disponível um conjunto limitado de dados ou um sistema de reconhecimento facial com elevado poder de discriminação.

Em [12] foi proposta a utilização de regularização aplicada à regressão logística como uma maneira de aliviar este problema, além de permitir obter valores mais conservadores (ou de menor magnitude, em escala logarítmica) para a LR, o que é desejado em situações em que poucos dados estão disponíveis para treinamento do modelo.

De forma simplificada, a regularização é obtida adicionando-se pseudo-dados com distribuições uniformes nos dois conjuntos de escores de treinamento. Para cada escore dos conjuntos de treinamento, são adicionados dois novos escores com o mesmo valor, um deles atribuído a H_p, mas com probabilidade $P = 0$, e o outro e H_d, mas com probabilidade

$P = 1$. Esses pseudo-escores adicionais são considerados, no processo de treinamento da regressão logística, com um peso reduzido em relação aos escores originalmente presentes no conjunto de treinamento, e a soma dos pesos atribuídos a esses pseudo-escores representa o fator de regularização aplicado. A Figura 4.9 ilustra o problema de separação completa dos conjuntos de treinamento e mostra dois exemplos de regularização, utilizando fatores de regularização distintos.

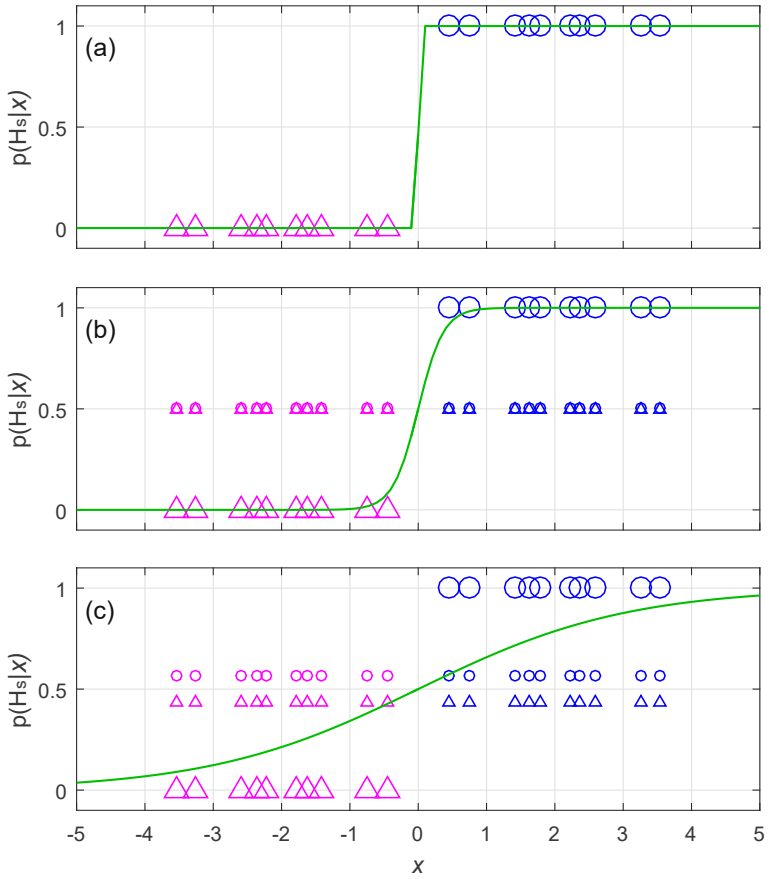


Figura 4.9: (a) Exemplo de regressão logística sem regularização em conjunto de dados com separação total dos escores. (b) Regressão logística com regularização, com baixo fator de regularização. (c) Regressão logística com regularização, com alto fator de regularização. Extraída de [12], com permissão.

Em relação ao efeito de limitação dos valores de LR obtidos pelo sistema, fatores mais altos de regularização induzem a uma limitação mais intensa das LR obtidas.

Embora a escolha da intensidade de regularização seja uma escolha arbitrária e, portanto, possível motivo de críticas a esta abordagem, a regressão logística regularizada tem vantagens em relação ao ELUB, principalmente por não haver cortes abruptos na conversão de escore para LR, o que evita o chamado *cliff-edge effect* discutido na seção anterior. Além disso, por ser também um método de cálculo de LR em si, não há estágios adicionais no sistema para limitação explícita do valor da LR.

4.4 Agregação

Esta seção descreve experimentos que exploram diferentes estratégias de agregação de *embeddings* aplicadas a cenários forenses, com a validação de sistemas de cálculo de LR empregando um estágio de agregação antes do cálculo do escore. A Figura 4.10 ilustra a abordagem de agregação de *embeddings* em comparação com o cenário tradicional, em que usualmente apenas a imagem de melhor qualidade é selecionada para cálculo do escore de similaridade e, posteriormente, da LR.

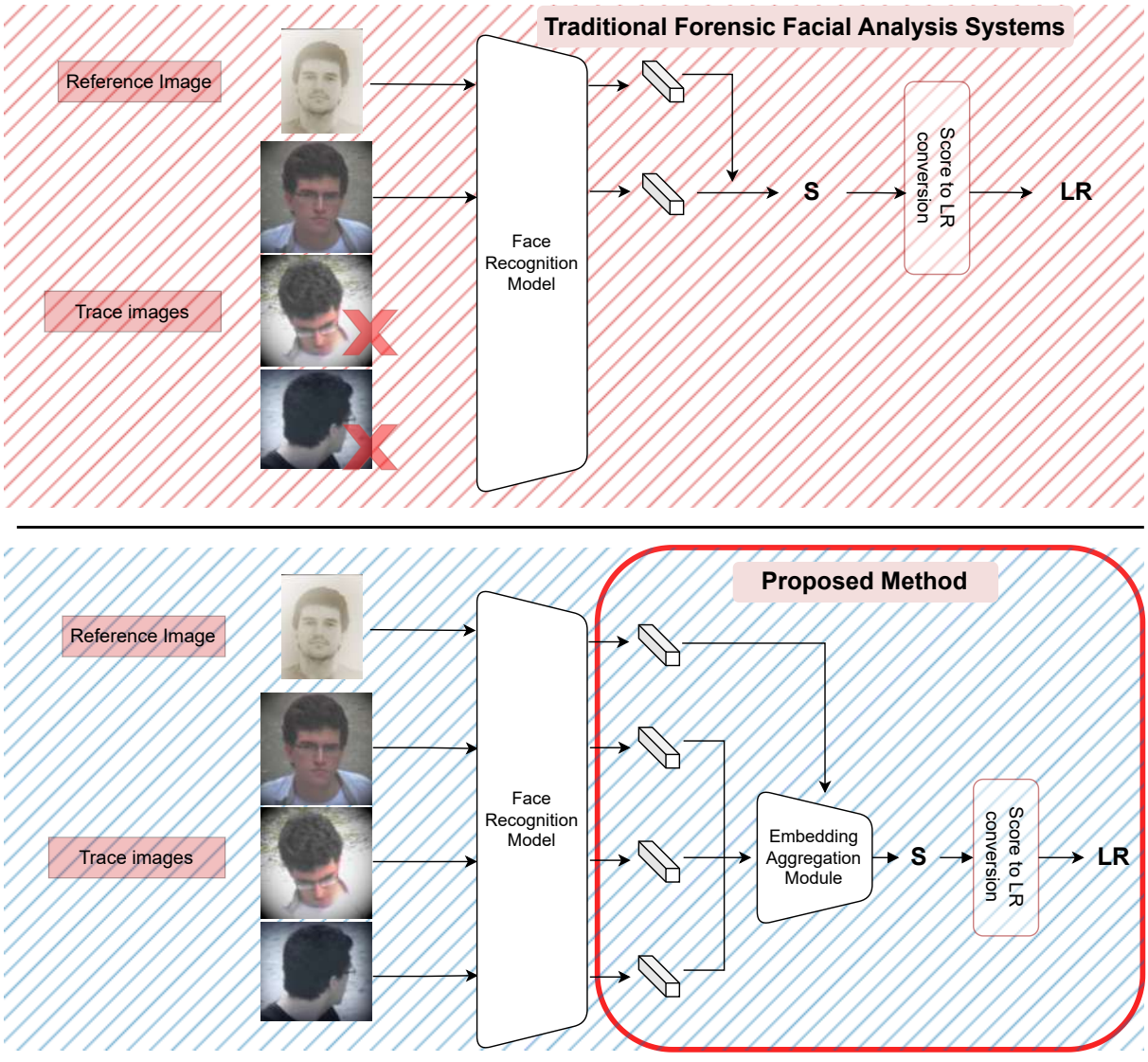


Figura 4.10: Abordagem proposta para agregação de *embeddings* em comparação à abordagem tradicional em que apenas a imagem de melhor qualidade é utilizada para cálculo da LR.

Da mesma forma como descrito na Seção 4.3, a LR é obtida a partir da calibração do escore de similaridade obtido da comparação entre imagens questionadas e padrões. O

método descrito nesta seção refere-se ao cálculo do escore, podendo a calibração do escore em LR seguir qualquer das estratégias descritas naquela seção. Embora os experimentos relacionados a agregação de *embeddings* realizados neste trabalho tenham sido focados no modelo de reconhecimento facial baseado em Arcface [54], resultados preliminares na base SCFace com outros dois modelos (FaceNet [24] e Dlib [102]) indicam que a agregação de *embeddings* pode oferecer melhora no desempenho para outros modelos de reconhecimento, conforme Tabela 4.1.

Distância / Modelo	1	2	3
Arcface baseline	0,367	0,057	0.007
Arcface AvgPool[4]	0,201	0,014	4,5 x 10^{-5}
FaceNet baseline	0,873	0,669	0.230
FaceNet AvgPool	0,779	0,502	0,133
Dlib baseline	0,930	0,733	0.641
Dlib AvgPool	0,866	0,645	0,552

Tabela 4.1: Melhora no desempenho de sistemas de cálculo de LR na base SCFace ao se utilizar agregação de *embeddings* de diversos modelos de reconhecimento facial.

Nos demais experimentos relacionados a agregação de *embeddings*, foi utilizado apenas um dos métodos de calibração do escore em LR: a regressão logística regularizada, sempre com fator inverso de regularização[5] igual a 1. Este método de calibração foi escolhido por não assumir uma distribuição específica dos escores de treinamento [11] e por ser menos sensível a variabilidade devido a amostragem [85, 30]. A escolha do fator inverso de regularização foi feita empiricamente, considerando uma relação de compromisso entre a limitação dos valores das LR obtidas e a manutenção do poder de discriminação do método.

4.4.1 Estratégias de agregação de *embeddings*

Os escores de similaridade obtidos a partir de *embeddings* do modelo ArcFace são calculados conforme a Equação 4.1:

$$s(\boldsymbol{t}, \boldsymbol{s}) = \frac{\boldsymbol{t} \cdot \boldsymbol{s}}{||\boldsymbol{t}||\,||\boldsymbol{s}||}$$

As estratégias investigadas nesta seção são empregadas para obtenção de uma única representação ($\boldsymbol{t}$, na Eq. 4.1) que combina as *embeddings* obtidas de cada uma das ima-

[4]AvgPool é uma das estratégias de agregação discutidas neste capítulo. Nesta estratégia a agregação é realizada pela média aritmética de cada componente das *embeddinds* a serem agregadas.

[5]parâmetro C na biblioteca scikit-learn [103].

gens questionadas que são atribuídas a um mesmo indivíduo de interesse. As estratégias investigadas podem ser descritas como uma combinação linear das N *embeddings*:

$$\boldsymbol{t} = \sum_{i=1}^{N} w_i \boldsymbol{t_i}, \tag{4.5}$$

em que $\boldsymbol{t_i}$ representa as *embeddings* de cada imagem questionada e w_i representa o peso a ser atribuído a cada imagem questionada. As estratégias de agregação avaliadas se diferenciam quanto à maneira de obtenção dos pesos w_i.

***Ser-Fiq* Pooling**

Com base na intuição de que imagens de melhor qualidade devem ter maior peso relativo na agregação das *embeddings*, foi empregado um modelo de estimação de qualidade da imagem considerado no estado da arte, proposto em [92], como critério para determinação dos pesos. Assim, os pesos w_i são obtidos a partir dos os escores de qualidade Ser-Fiq normalizados s_i de cada imagem questionada:

$$w_i = \frac{s_i}{\sum_{j=1}^{N} s_j}, \tag{4.6}$$

CS Pooling

Também foi considerado outro método, proposto em [93], para estimação de qualidade de imagens faciais. Nesta estratégia, os pesos w_i são determinados a partir dos *confusion scores* cs_i pela seguinte equação:

$$w_i = \frac{1 - cs_i}{\sum_{j=1}^{N}(1 - cs_j)}, \tag{4.7}$$

pois imagens de melhor qualidade resultam em *confusion scores* mais baixos.

Average Pooling

Nesta estratégia, são atribuídos pesos iguais a todas as imagens questionadas, sendo equivalente à obtenção de uma média aritmética simples das componentes das *embeddings*.

4.4.2 Estratégias de agregação de escores

Além das estratégias de agregação de *embeddings*, também foram avaliadas outras duas estratégias que permitem agregar informações de um conjunto de imagens questionadas. Estas duas estratégias se baseiam na agregação de escores, obtidos a partir das comparações possíveis entre a imagem padrão e cada imagem questionada. Assim, foram conside-

radas as estratégias *MaxScore*, em que o maior escore desse conjunto de comparações é utilizado para cálculo da LR, e *AvgScore*, em que a média dos escores é considerada como o escore da comparação entre a imagem padrão e o conjunto de imagens questionadas.

Embora estas estratégias não se refiram à agregação de *embeddings*, elas foram consideradas por serem facilmente implementadas na prática em todos os cenários em que as estratégias de agregação de *embeddings* também são aplicáveis.

4.4.3 Novo protocolo para verificação na base Quis-Campi

Com o objetivo de avaliar a agregação de *embeddings* de forma mais realista na base Quis-Campi, é proposto um novo protocolo baseado no conceito de *encontros*. Neste protocolo, as imagens de videomonitoramento de cada identidade são agrupadas em conjuntos capturados a cada encontro com o sistema de captura de imagens. Para este fim, foi escolhido um limiar de dois minutos entre cada captura como critério para separação dos encontros. A revisão manual por amostragem dos agrupamentos determinados com base neste critério indicou que ele foi adequado para separar capturas em cenas distintas. Sob este protocolo, apenas as *embeddings* das imagens de um mesmo encontro são agregradas para cálculo de escore de similaridade. Este protocolo é representativo de casos onde imagens de uma pessoa de interesse são registradas no vídeo e nenhuma outra imagem de videomonitoramento pode ser seguramente atribuída ao mesmo indivíduo, ainda que ele apareça em imagens capturadas em outros momentos. Os resultados obtidos sob este protocolo são referenciados como *Quis-Campi encounters* e os metadados indicando as imagens atribuídas a cada encontro são disponibilizados no endereço `https://github.com/rafribeiro/embedding_aggregation`.

4.5 Validação

Para avaliar se a utilização de um determinado sistema de cálculo de LR é recomendada em um determinado caso, é preciso validar o sistema nas condições específicas do caso. Alternativamente, é possível validar previamente o sistema em condições que, pela experiência do perito, retratem um determinado conjunto de casos com características comuns entre si.

Essa validação procura **estabelecer se o sistema de cálculo de LR fornece informação que melhora a qualidade das decisões** da instância julgadora, destinatária e usuária da evidência produzida no exame pericial.

Os critérios de validação utilizados nessa pesquisa e que serão apresentados nesta seção estão baseados, principalmente, em dois trabalhos. No primeiro, Meuwly *et al.* [31] apresentam diretrizes gerais para a validação de sistemas de cálculo de LR a partir de

escores. No segundo, Morrison *et al.* [13] elaboraram um consenso para validação de sistemas forenses de comparação de locutores. Este segundo trabalho, embora enfoque na biometria da voz/fala, apresenta métricas e representações gráficas de desempenho que são perfeitamente utilizáveis por outros tipos de sistemas biométricos.

Para obtenção das LRs de validação de cada sistema, os seguintes procedimentos são adotados. As LRs são obtidas através de uma estratégia de validação cruzada do tipo *Leave-One-Person-Out* e *Leave-Two-Persons-Out*, para os conjuntos de LRs relacionadas a H_p e a H_d, respectivamente[6]. Assim, para obter as LRs relacionadas a H_p, a cada rodada da validação cruzada, as imagens de uma pessoa presente na base são separadas e as imagens das demais pessoas são utilizadas para treinar o sistema de conversão de escore em LR[7]. Os escores obtidos entre as imagens da pessoa que foi separada são então convertidos em LRs utilizando o sistema treinado naquela rodada, e as LRs obtidas são adicionadas ao conjunto de LRs relacionadas a H_p. Já para obter as LRs relacionadas a H_d, a cada rodada da validação cruzada é necessário separar imagens de duas pessoas distintas e realizar o treinamento do sistema de conversão de escore para LR a partir dos escores obtidos entre as imagens das demais pessoas. Os escores obtidos das comparações entre as imagens das duas pessoas separadas são então convertidos em LRs utilizando o sistema treinado naquela rodada, e essas LRs são adicionadas ao conjunto de LRs relacionadas a H_d. Ao final de todas as todas de validação cruzada, considera-se os conjuntos de LRs relacionadas a H_p e a H_d para obtenção das métricas e representações gráficas para validação do sistema.

4.5.1 *Log Likelihood Ratio Cost* - C_{llr}

O custo do logaritmo da LR, ou C_{llr}, é uma medida que permite avaliar o poder de discriminação e o grau de calibração de um sistema de cálculo de LR. [84, 97]

É uma medida expressa através de um único valor escalar e pode ser calculada pela fórmula apresentada na Equação 3.1, repetida a seguir para conveniência do leitor:

$$C_{\text{llr}} = \frac{1}{2}\left[\frac{1}{N_p}\sum_{i_p} log_2\left(1+\frac{1}{LR_i}\right) + \frac{1}{N_d}\sum_{j_d} log_2(1+LR_j)\right]$$

A Figura 4.11, adaptada de [13], ilustra o comportamento dos termos dentro de cada somatório na equação que define a C_{llr}. Visualmente, é possível verificar que erros de

[6]Nos casos das bases Quis-Campi, BFW e BFW clean, foi utilizada estratégia de validação cruzada do tipo *k-fold*, com $k = 100$, como forma de aliviar o custo computacional.

[7]No caso de métodos paramétricos ou do método KDE, esse treinamento consiste em estimar as funções densidade de probabilidade relacionadas a H_p e a H_d. No caso da regressão logística, consiste em treinar o modelo de regressão.

magnitude mais elevada são penalizados de maneira mais severa do que erros de menor magnitude (mais próximos do valor neutro da LR). Assim, a C_{llr} pode ser inicialmente compreendida como uma taxa de erros ponderada pela magnitude dos erros.

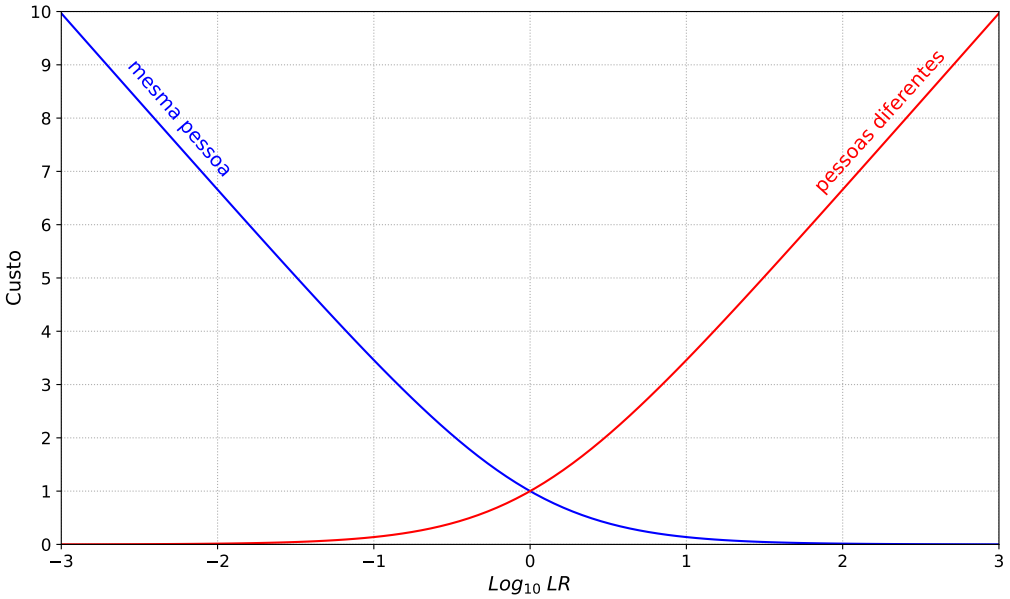


Figura 4.11: Funções de custo que compõem a C_{llr}. Adaptada de [13].

A C_{llr} apresenta diversas propriedades interessantes para a avaliação de desempenho de um sistema de cálculo de LR. Primeiro, é uma métrica que é independente de cada caso em que o sistema será utilizado, onde diferentes custos/utilidades e razões de probabilidade *a priori* podem ser aplicadas a cada caso. Em segundo lugar, por ser representada por um único valor escalar, é fácil utilizar a C_{llr} para comparar diferentes sistemas de cálculo de LR. Terceiro, a C_{llr} tem uma interpretação baseada em teoria da informação: dado um sistema que calcula LRs, o valor $1 - C_{\text{llr}}$ representa a quantidade de informação entregue pelo sistema ao usuário (neste caso, à instância julgadora), assumindo máxima entropia *a priori*, ou seja, $P(H_p) = P(H_d) = \frac{1}{2}$. Assim, quanto menor o valor da C_{llr}, maior é a informação entregue pelo sistema à instância julgadora. Finalmente, a C_{llr} pode ser decomposta em duas componentes: um custo (ou perda) de discriminação ($C_{\text{llr}}{}^{min}$) e um custo associado a calibração ($C_{\text{llr}}{}^{cal}$). Essas duas componentes podem ser obtidas através de uma calibração ótima do conjunto de validação, através de uma transformação monotônica das LRs desse conjunto [84].

Como visto, é desejável que os sistemas para cálculo de LR possuam C_{llr} baixa. Assim, a partir da interpretação da C_{llr} baseada em teoria da informação, pode-se estabelecer como critério de validação que sistemas com $C_{\text{llr}} < 1$ são apropriados para uso forense,

desde que apresentem boa calibração. Sistemas que atendem esse critério contribuem para reduzir a incerteza das decisões da instância julgadora, diminuindo a entropia dessas decisões. De fato, este mesmo critério é recomendado por Morrison *et al.* em [13].

A avaliação quanto à calibração pode ser realizada tanto em termos quantitativos, através da componente da C_{llr} relativa à calibração ($C_{\text{llr}}{}^{cal}$), quanto por inspeção dos gráficos Tippett.

4.5.2 Gráficos Tippett

Gráficos Tippett apresentam distribuições de probabilidades acumuladas empíricas de LR. É baseado em análises feitas inicialmente em 1968 por Tippett et al [81] e proposto posteriormente por Meuwly em [104] no formato utilizado atualmente, mostrando no mesmo gráfico as curvas de distribuição de probabilidade acumulada empírica das LRs associadas a cada uma das hipóteses.

Neste tipo de gráfico, quanto maior o afastamento entre as curvas na direção horizontal, melhor é o desempenho do sistema em termos de poder de discriminação. E quanto mais simetricamente distribuídas as curvas se encontrarem em relação ao valor neutro da LR, melhor é o desempenho do sistema em termos de calibração.

Uma vantagem do gráfico Tippett é que, por ser um gráfico que apresenta dados acumulados empíricos do sistema sob avaliação, as curvas representam exatamente os dados obtidos pelo sistema, diferentemente de outros tipos de gráficos, como histogramas ou curvas obtidas por estimativa de densidade de kernel.

A Figura 3.7 mostrou um exemplo de gráfico Tippett, mas nesta pesquisa será adotada uma versão modificada, na qual a curva referente às LR associadas à hipótese de mesma origem (H_p) tem significado invertido, ou seja, ela deve ser lida como "proporção de LRs cuja hipótese H_p é verdadeira que são menores que". Essa inversão da curva relacionada a H_p, no entendimento do autor, facilita a visualização de algumas propriedades importantes do sistema sob análise, especialmente quanto à calibração. A Figura 4.12, adaptada de [13] ilustra alguns exemplos de gráficos Tippett construídos segundo essa interpretação.

O ponto central no eixo horizontal corresponde ao valor neutro da LR ($log_{10}\, LR = 0$) e, utilizando gráficos Tippett com a interpretação indicada no parágrafo anterior, é possível facilmente perceber problemas de calibração, evidenciada pelo desvio das curvas para a direita ou para a esquerda deste ponto central. Também é possível avaliar duas métricas secundárias relacionadas ao desempenho de sistemas para fins forenses: i. RMEp - do inglês *Rate of Misleading Evidence in favor of the prosecution*; e ii. RMEd - do inglês *Rate of Misleading Evidence in favor of the defense.*

Essas duas métricas podem ser obtidas pelo valor de cada curva no ponto central e representam a proporção de casos em que era verdadeira a hipótese da **defesa** e a LR foi

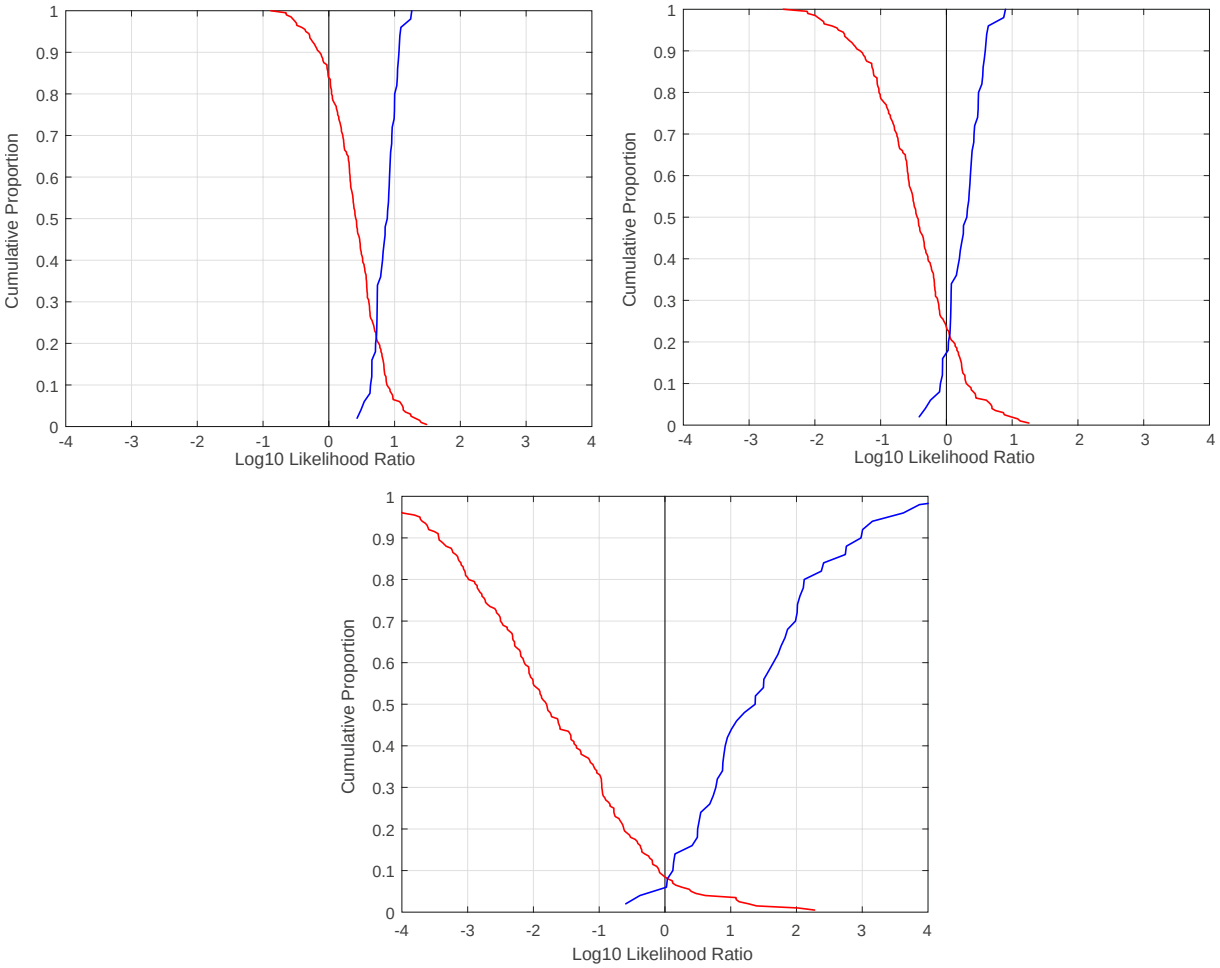


Figura 4.12: Gráficos Tippett de três sistemas distintos. No topo à esquerda, um sistema mal calibrado e com baixo poder de discriminação. No topo à direita, o mesmo sistema após calibração. Na parte inferior, um sistema bem calibrado, com maior poder de discriminação. Adaptada de [13], com permissão.

maior do que 1, e a proporção de casos em que era verdadeira a hipótese da **acusação** e a LR foi **menor** do que 1, respectivamente.

Como segundo critério de validação, os gráficos Tippett devem indicar boa calibração, ou seja, não pode haver desvios sistemáticos óbvios em relação ao valor neutro da LR.

Capítulo 5

Resultados

Este capítulo apresenta os resultados obtidos na pesquisa, divididos em duas seções. Na Seção 5.1 são apresentados os resultados dos experimentos de calibração de escore em LR, empregando os modelos FaceNet e ArcFace, sem agregação de *embeddings*, segundo os diferentes métodos de calibração avaliados: gaussiana normal, KDE, regressão logística e regressão logística com regularização. Na Seção 5.2 são apresentados resultados dos experimentos relacionados a agregação de *embeddings* e de escores.

5.1 Resultados sem agregação de *embeddings*

A Tabela 5.1 mostra os valores de C_{llr} obtidos para cada base em função do modelo de reconhecimento (ArcFace ou FaceNet) e do método de calibração de escore em LR. (RL = regressão logística; RLR = regressão logística regularizada; KDE = *kernel density estimation*; Gaussian = gaussiana normal).

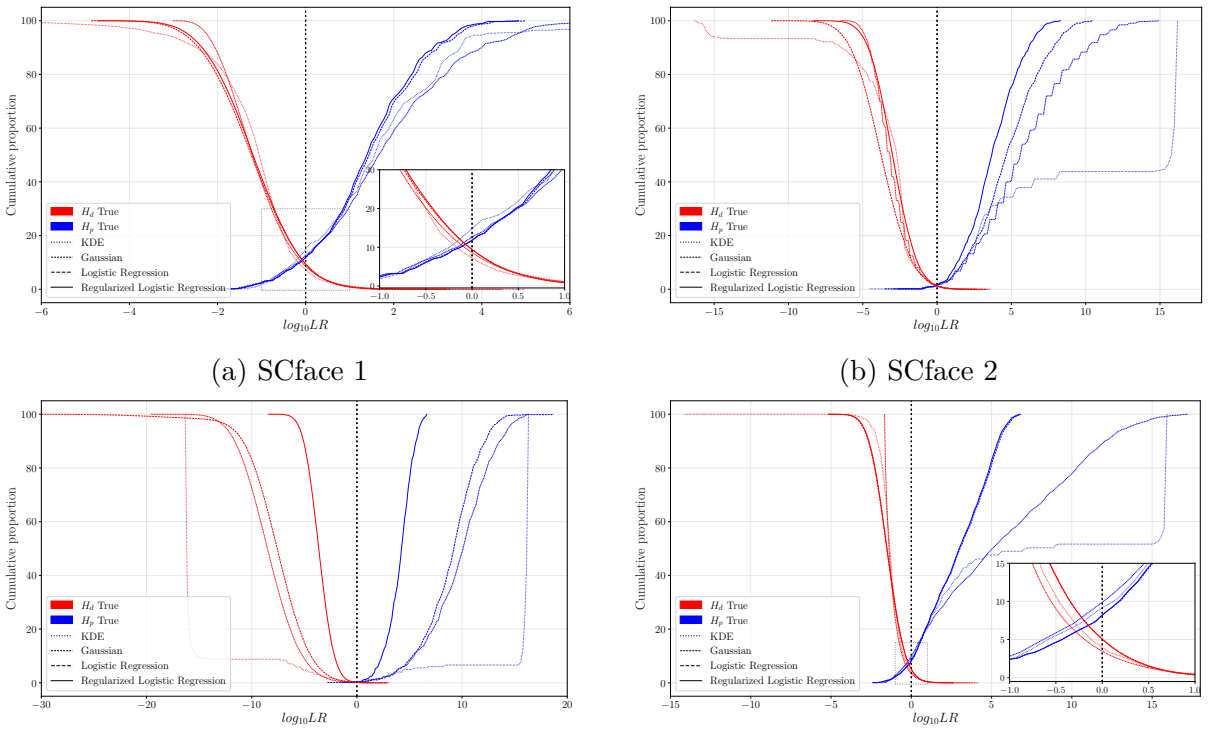
Tabela 5.1: C_{llr} para os diversos métodos de calibração aplicados às bases utilizadas.

Base Método	SCface 1	SCface 2	SCface 3	SCface all	Quis-Campi	BFW	BFW clean	FEI
ArcFace								
RL	0,367	0,059	0,013	0,249	0,226	0,217	0,083	$\mathbf{2x10^{-8}}$
RLR	0,367	0,060	0,011	0,249	0,226	0,217	0,083	$1x10^{-4}$
KDE	0,367	0,062	**0,008**	**0,245**	**0,215**	**0,211**	**0,079**	$2x10^{-4}$
Gaussian	**0,366**	**0,057**	0,007	0,252	0,229	0,245	0,111	$8x10^{-8}$
FaceNet								
RL	0,904	0,672	0,287	0,684	0,624	0,302	0,242	0,050
RLR	0,904	0,672	0,287	0,684	0,624	0,302	0,242	0,050
KDE	0,898	0,665	0,279	0,668	0,613	0,301	0,240	0,050
Gaussian	0,897	0,663	0,294	0,670	0,611	0,321	0,266	0,072

Todas as combinações de sistemas de reconhecimento facial e método de calibração de escore em LR satisfazem o critério de que a C_{llr} deve ser menor do que 1. No caso de algumas das combinações envolvendo o FaceNet, porém, a C_{llr} é bastante próxima desse limite (SCface 1) e, em todos os casos, o método correspondente envolvendo o ArcFace apresentou desempenho superior (menor C_{llr}), o que é coerente com o maior poder de discriminação do ArcFace relativamente ao FaceNet.

Em relação aos métodos de calibração, observa-se pouca variação dos valores C_{llr} para cada combinação de base e modelo de reconhecimento, exceto nos valores obtidos pelo método Gaussian nas bases BFW e BFW clean, que foram relativamente piores do que os demais, tanto no caso do ArcFace quanto no caso do FaceNet. O pior desempenho desse método nas duas bases é atribuído ao desvio da condição de normalidade nas distribuições dos escores relacionados a H_p, conforme se verifica na Figura 4.5. Uma avaliação geral sobre o desempenho dos métodos em cada base será feita após a apresentação e discussão dos gráficos Tippett.

Considerando a expressiva diferença entre os valores de C_{llr} entre os métodos baseados no ArcFace em relação ao FaceNet, serão exibidos e discutidos apenas os gráficos Tippett relacionados ao ArcFace. A Figura 5.1 mostra os gráficos Tippett para os subconjuntos da base SCface.



(a) SCface 1

(b) SCface 2

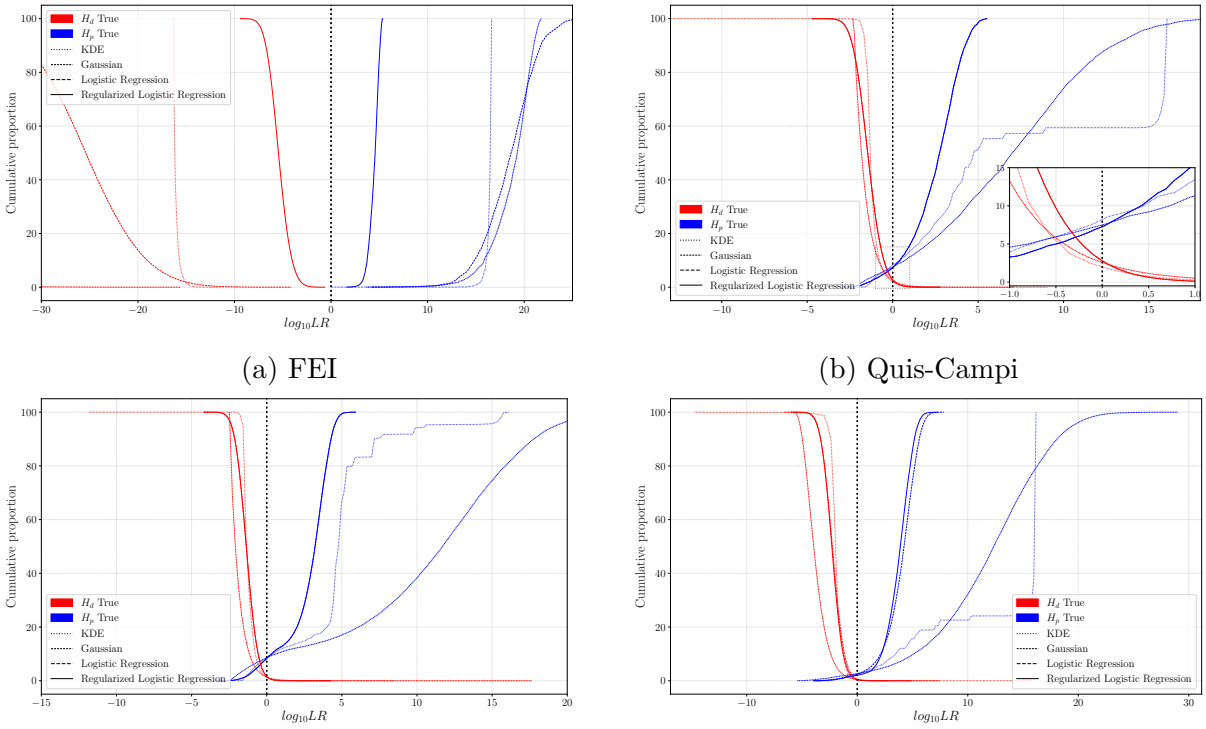
(c) SCface 3

(d) SCface all

Figura 5.1: Gráficos Tippett para os subconjuntos da base SCface. O retângulo em destaque em alguns gráficos mostra detalhes em torno de $log_{10}\, LR = 0$.

Observa-se nesta Figura que apesar dos métodos KDE e Gaussian terem apresentado melhor C_{llr} do que os baseados em regressão logística, esses métodos (KDE e Gaussian) geram LRs de magnitude muito elevadas, difíceis de serem justificadas a partir de um modelo de calibração obtido a partir de uma base com apenas 130 identidades. Este mesmo fenômeno é observado para a regressão logística (sem regularização) no subconjunto SC-face 3, em que há pouca sobreposição entre os conjuntos de escores.

A Figura 5.2 mostra os gráficos Tippett das demais bases: FEI, BFW, BFW clean e Quis-Campi.



(a) FEI

(b) Quis-Campi

(c) BFW

(d) BFW clean

Figura 5.2: Gráficos Tippett para as bases FEI, Quis-Campi, BFW e BFW clean.

O mesmo efeito de obtenção de LRs de valores extremos é observado para os métodos KDE e Gaussian nestas outras bases. Na base FEI, particularmente, a maior parte da curva do método Gaussian não pôde ser exibida para evitar que as demais curvas ficassem com a visualização prejudicada. Mesmo considerando o elevado poder de discriminação do ArcFace e a ótima qualidade das imagens da base FEI, muitos valores de LR obtidos nessa base para os métodos KDE, Gaussian e Regressão Logística (sem regularização) representariam evidências mais fortes do que as apresentadas em exames de DNA, o que é, novamente, difícil de justificar considerando o tamanho dessa base.

Assim, ainda que todos os métodos baseados no ArcFace tenham atendido aos requisitos apresentados na literatura para serem considerados como adequados para fins forenses,

a geração de valores extremos de LR especialmente nos métodos KDE e Gaussian sugerem que estes métodos devem ser evitados. No caso do método Gaussian, observa-se que não há, na prática, qualquer vantagem na sua utilização em relação à regressão logística (com ou sem regularização), pois no caso em que as distribuições de escores sejam perfeitamente normais, a regressão logística produz os mesmos resultados do método Gaussian [11].

Dentre os métodos avaliados, a Regressão Logística Regularizada apresenta portanto um melhor equilíbrio entre a manutenção do poder de discriminação (apesar da regularização) e a não geração de valores extremamente elevados para as LR. Além disso, o fato deste método não assumir uma distribuição específica dos escores de treinamento é uma vantagem em relação aos métodos paramétricos. Em relação ao KDE, os métodos baseados em regressão logística apresentam ainda a vantagem de garantia de monotonicidade na conversão de escore para LR.

5.2 Resultados com agregação

A fim de avaliar se e quanto a agregação de *embeddings* melhora o desempenho de sistemas de cálculo de LR, foram utilizados como referência de desempenho (*baseline*) o método convencional de calibração de escore em LR, considerando cada imagem questionada independentemente, sendo obtida uma LR para cada comparação entre a imagem de referência e as imagens questionadas. Conforme descrito no início deste capítulo, apenas a regressão logística regularizada foi utilizada como método de calibração, e apenas o ArcFace foi usado como modelo de reconhecimento facial. A distribuição do número de *embeddings* que são agregadas por identidade em cada base é mostrada na Figura 5.3.

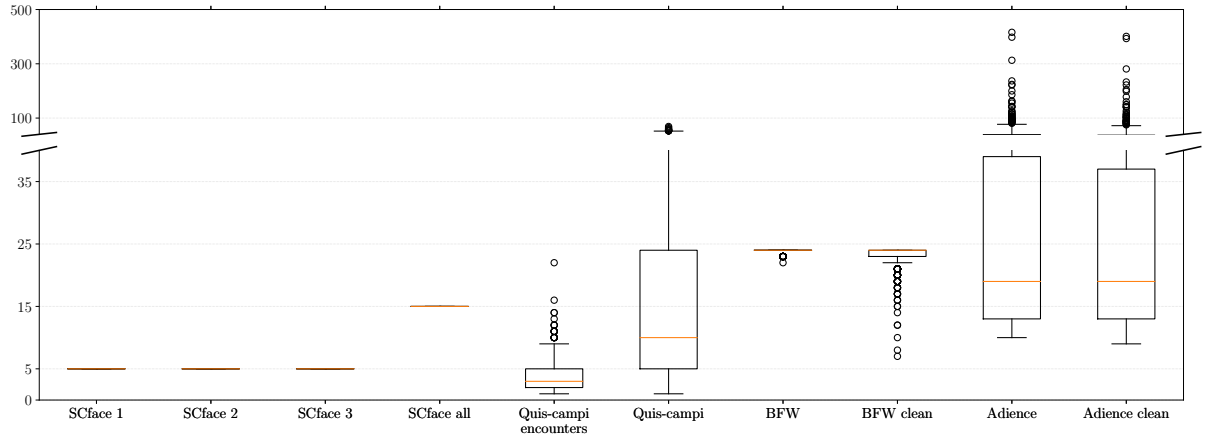


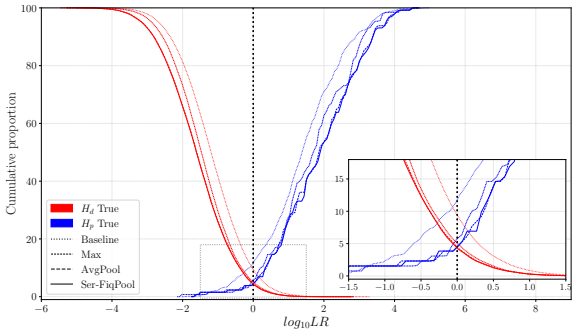
Figura 5.3: Distribuição do número de *embeddings* agregadas por idendidade em cada base.

Os resultados para as bases SCface e Quis-Campi são mostrados na Tabela 5.2 e na Figura 5.4. Primeiramente, observa-se uma melhoria substancial na C_{llr} em relação a Mandasari et al. [82] na base SCFace [5]. Essa melhoria é atribuída primordialmente ao poder discriminatório do modelo de reconhecimento facial, uma vez que mesmo a abordagem *baseline*, sem agregação, ofereceu resultados substancialmente melhores. Em relação às abordagens de agregação de *embeddings*, observa-se que aquelas baseadas em ponderação por qualidade (CSPool e Ser-FiqPool) apresentaram melhor desempenho, mas apenas marginalmente melhor do que a abordagem AvgPool. Além disso, as melhorias mais significativas em C_{llr} ocorreram para as bases que dispunham de mais *embeddings* para serem agregadas (SCface all e Quis-Campi). Também foi observado uma melhora substancial a partir da agregação de *embeddings* de imagens com resoluções mais baixas (SCface 1 e SCface 2). Em suma, verifica-se que as estratégias de agregação são especialmente interessantes em cenários com imagens de pior qualidade (SCface 1) e com maiores quantidades de imagens de um mesmo indivíduo (SCface all e Quis-Campi).

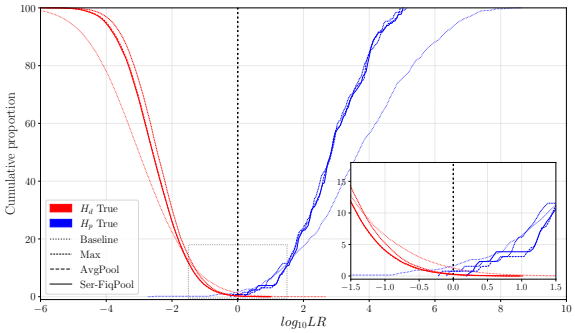
Tabela 5.2: C_{llr} para as bases SCface e Quis-Campi

Base / Método	SCface 1	SCface 2	SCface 3	SCface all	Quis-Campi encounters	Quis-Campi
[82] Raw scores	0.659	0.313	0.378	0.503	-	-
[82] ZT-norm scores	0.664	0.243	0.287	0.419	-	-
Baseline	0.368	0.060	0.011	0.249	0.226	0.226
AvgScore	0.221	0.037	0.013	0.023	0.209	0.105
MaxScore	0.234	0.035	0.011	**0.012**	0.222	0.115
AvgPool	0.212	0.029	**0.010**	0.013	0.202	0.098
CSPool	0.210	0.029	**0.010**	**0.012**	0.201	0.098
Ser-FiqPool	**0.209**	**0.028**	**0.010**	0.018	**0.198**	**0.095**

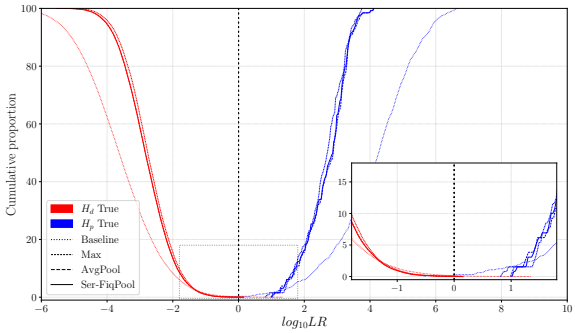
Os resultados para o cenário de redes sociais (bases Adience e BFW, e suas derivadas) são apresentados na Tabela 5.3 e na Figura 5.5. Primeiramente, observa-se uma melhora significativa no desempenho após a limpeza das duas bases. Também foram observados ganhos nas estratégias de agregação avaliadas em relação ao método *baseline*. Para algumas bases, a estratégia *MaxScore* apresentou desempenho superior (menor C_{llr} e maior separação dos gráficos Tippett) em relação às estratégias de agregação de *embeddings*. Especula-se que isso se deve à presença de imagens capturadas na mesma sessão (por exemplo, quadros consecutivos de uma gravação de vídeo), que tendem a produzir escores muito altos quando uma dessas imagens é selecionada como referência, mas oferecem informações redundantes para as estratégias que agregam várias *embeddings*. Esta possibilidade não foi investigada em razão do foco destes experimentos estar na agregação de *embeddings* e não de escores.



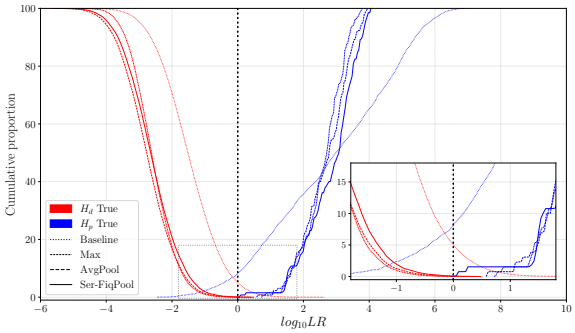
(a) SCface 1



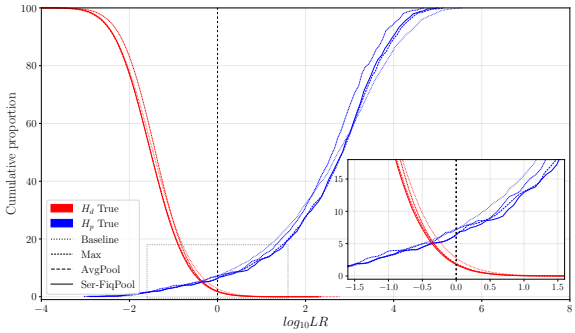
(b) SCface 2



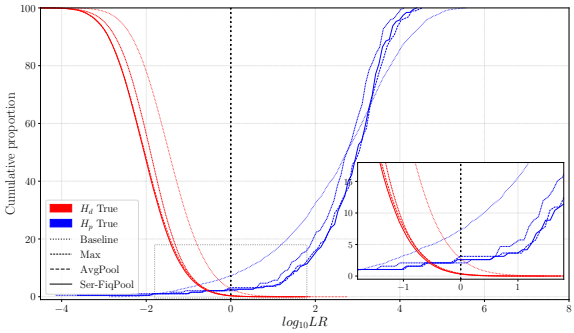
(c) SCface 3



(d) SCface all



(e) Quis-Campi encounters



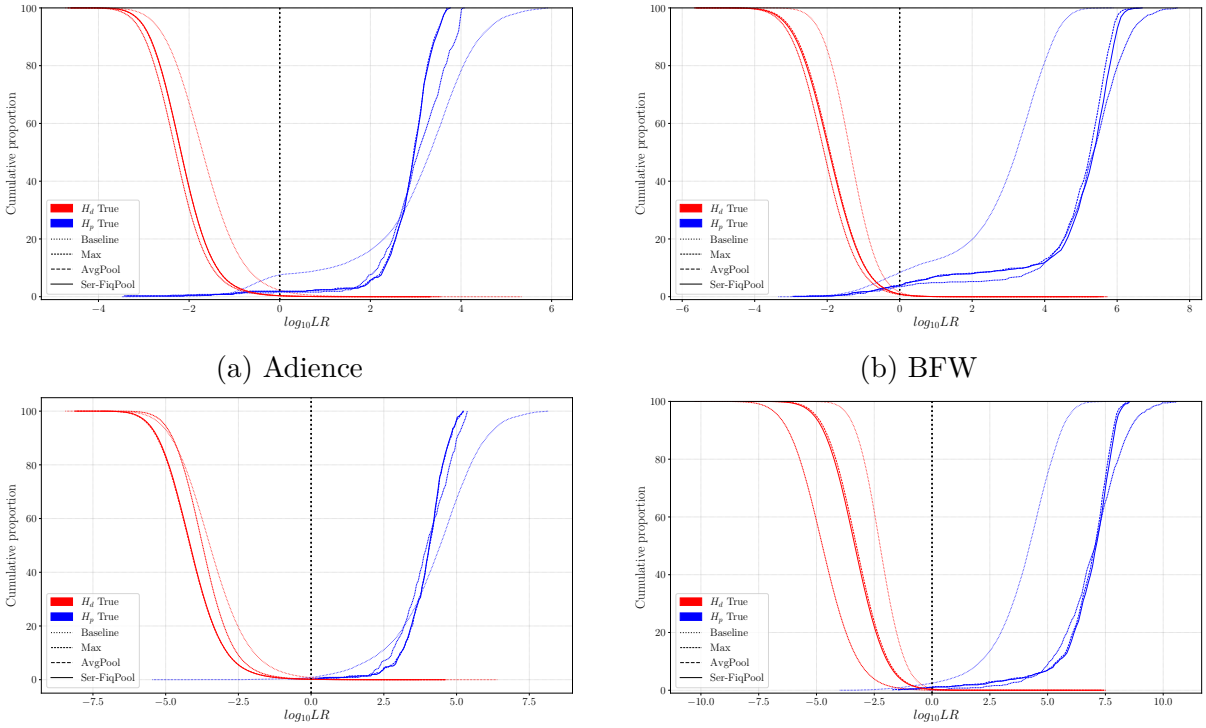
(f) Quis-Campi

Figura 5.4: Gráficos Tippett para as bases do cenário de videomonitoramento. O retângulo em destaque mostra detalhes de cada gráfico em torno de $log_{10} LR = 0$. Curvas Tippett para algumas estratégias foram omitidas para não sobrecarregar a visualização.

Tabela 5.3: C_{llr} para o cenário de redes sociais

Base / Método	Adience	Adience clean	BFW	BFW clean
Baseline	0.174	0.038	0.217	0.083
AvgScore	0.069	0.008	0.129	0.036
MaxScore	**0.058**	0.010	**0.088**	**0.003**
AvgPool	0.068	0.007	0.114	0.027
CSPool	0.068	**0.006**	0.114	0.026
Ser-FiqPool	0.068	**0.006**	0.112	0.025

Os resultados dos experimentos apresentados nesta seção demonstram que agregar *embeddings* de várias imagens da mesma pessoa é uma técnica eficaz para melhorar o desempenho do reconhecimento facial, principalmente para imagens de baixa resolução, o que é especialmente relevante para as condições mais desafiadoras em casos forenses. Até mesmo abordagens mais simples, como *AvgPool*, podem oferecer melhorias substanciais de desempenho quando as imagens são de qualidade similar. Esta estratégia também tem a vantagem de não exigir a estimativa da qualidade da imagem facial, sendo mais barata computacionalmente.



(a) Adience

(b) BFW

(c) Adience clean

(d) BFW clean

Figura 5.5: Gráficos Tippett para as bases do cenário de redes sociais. Curvas Tippett para algumas estratégias foram omitidas para não sobrecarregar a visualização.

Capítulo 6

Conclusão

Esta pesquisa avaliou diferentes métodos de calibração de escores obtidos de sistemas de reconhecimento facial em LR. Os métodos foram avaliados em cenários tipicamente encontrados em exames periciais: imagens de videomonitoramento e imagens de mídias sociais.

Os critérios utilizados para avaliar os métodos foram adaptados da literatura da área de reconhecimento automático de locutor, especialmente do consenso publicado em 2021 [13].

Foi observado que mesmo em cenários considerados desafiadores para reconhecimento facial, os sistemas biométricos no estado da arte possuem desempenho suficiente para que sua utilização seja recomendada como ferramenta de análise forense, dando suporte à produção de prova material mais robusta. Além disso, métodos para cálculo de LR baseados em sistemas biométricos possuem uma vantagem relevante em relação aos métodos tradicionais, baseados em análises manuais e qualitativas da morfologia facial: o desempenho desses métodos pode ser validado empiricamente nas condições específicas de cada caso, seguindo os mesmos procedimentos realizados nesta pesquisa.

Também foram avaliadas estratégias de agregação de *embeddings* aplicadas a cenários forenses. Os experimentos relacionados a essas estratégias demonstraram que é viável obter desempenho ainda melhor dos modelos de reconhecimento facial já disponíveis em situações onde estejam disponíveis múltiplas imagens do mesmo indivíduo, principalmente em cenários em que as imagens sejam de baixa qualidade.

Este trabalho ofereceu ainda as seguintes contribuições não previstas originalmente no Projeto de Pesquisa:

1. um novo protocolo de verificação para a base Quis-Campi, mais representativo de situações efetivamente encontradas em exames periciais;

2. um modelo de reconhecimento com desempenho compatível com estado da arte em reconhecimento facial, validado para fins forenses e cuja licença permite o uso para fins periciais; e

3. versões limpas das bases Adience e BFW, com a eliminação de muitos erros nos rótulos de identidade e de imagens duplicadas.

Como limitações desta pesquisa, destaca-se primeira e principalmente a utilização de escores que consideram apenas a similaridade entre as imagens faciais, sem levar em conta a tipicidade das faces. [32] demonstrou através de simulações Monte-Carlo que escores que consideram apenas a similaridade não são ideais para o cálculo de LR. De fato, os sistemas de cálculo de LR empregados na área de comparação de locutor empregam escores que consideram a tipicidade das amostras de voz.

Outra limitação encontrada foi a ausência de imagens de referência nas bases do cenário de redes sociais, com pose, iluminação e expressão facial controladas. Esta é uma diferença importante em relação aos casos periciais envolvendo imagens questionadas obtidas em redes sociais. Além disso, especificamente para os experimentos relacionados a agregação de *embeddings* a presença de múltiplas imagens da mesma sessão nessas bases torna mais difícil generalizar os resultados das estratégias de agregação baseadas em escores (*AvgScore* e *MaxScore*) para cenários forenses.

Além disso, também é apontada como limitação desta pesquisa a não exploração de um cenário frequentemente encontrado em casos reais: imagens faciais impressas em documentos de identificação. Ciente deste cenário relevante, o autor iniciou a coleta de uma base de imagens neste cenário, entretanto não foi possível concluir a coleta a tempo de incluir este cenário nos experimentos, ficando a avaliação deste cenário como trabalho futuro.

Como principal trabalho futuro, pretende-se investigar a utilização de escores que levem em conta a similaridade e a tipicidade das imagens faciais. Além disso, pretende-se coletar base de imagens faciais em documentos impressos e realizar experimentos similares aos apresentados nesta pesquisa, bem como avaliar estratégias de agregação de *embeddings* mais sofisticadas, como [74, 105]. Por fim, a avaliação de diferentes modelos de reconhecimento facial, treinados com diferentes arquiteturas e utilizando diferentes bases de imagens faciais, é outro aspecto deixado como trabalho futuro.

Referências

[1] Phillips, P. Jonathon, Amy N. Yates, Ying Hu, Carina A. Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G. Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, Jun Cheng Chen, Carlos D. Castillo, Rama Chellappa, David White e Alice J. O'Toole: *Face recognition accuracy of forensic examiners, super-recognizers, and face recognition algorithms.* Proceedings of the National Academy of Sciences, 115(24):6171–6176, 2018, ISSN 0027-8424. `https://www.pnas.org/content/115/24/6171`. x, 3, 4

[2] Kanade, Takeo: *Computer recognition of human faces.* Birkhäuser Basel, 1977. `https://doi.org/10.1007/978-3-0348-5737-6`. x, 12, 14

[3] Taigman, Yaniv, Ming Yang, Marc'Aurelio Ranzato e Lior Wolf: *Deepface: Closing the gap to human-level performance in face verification.* Em *2014 IEEE Conference on Computer Vision and Pattern Recognition*, páginas 1701–1708, 2014. x, 14, 15

[4] Zeiler, Matthew D. e Rob Fergus: *Visualizing and understanding convolutional networks.* Em Fleet, David, Tomas Pajdla, Bernt Schiele e Tinne Tuytelaars (editores): *Computer Vision – ECCV 2014*, páginas 818–833, Cham, 2014. Springer International Publishing, ISBN 978-3-319-10590-1. x, 15

[5] Grgic, Mislav, Kresimir Delac e Sonja Grgic: *Scface — surveillance cameras face database.* Multimedia Tools Appl., 51(3):863–879, feb 2011, ISSN 1380-7501. `https://doi.org/10.1007/s11042-009-0417-2`. x, 17, 19, 59

[6] Gonzalez-Rodriguez, Joaquin, Julian Fierrez-Aguilar, Daniel Ramos-Castro e Javier Ortega-Garcia: *Bayesian analysis of fingerprint, face and signature evidences with automatic biometric systems.* Forensic Science International, 155(2):126–140, 2005, ISSN 0379-0738. `https://www.sciencedirect.com/science/article/pii/S0379073804007509`. x, 20, 21, 22, 23, 34

[7] Meuwly, D.: *Forensic individualisation from biometric data.* Science & Justice, 46(4):205–213, 2006, ISSN 1355-0306. `https://www.sciencedirect.com/science/article/pii/S1355030606716008`. x, 9, 22, 23, 26

[8] Jacquet, Maëlig e Christophe Champod: *Automated face recognition in forensic science: Review and perspectives.* Forensic Science International, 307:110124, 2020, ISSN 0379-0738. `https://www.sciencedirect.com/science/article/pii/S0379073819305365`. xi, xiii, 9, 20, 25, 26, 27, 28, 34

[9] Morrison, Geoffrey Stewart, Jonas Lindh e James M Curran: *Likelihood ratio calculation for a disputed-utterance analysis with limited available data*. Speech Communication, 58:81–90, 2014, ISSN 0167-6393. https://www.sciencedirect.com/science/article/pii/S0167639313001635. xi, 30

[10] Vergeer, Peter, Andrew van Es, Arent de Jongh, Ivo Alberink e Reinoud Stoel: *Numerical likelihood ratios outputted by lr systems are often based on extrapolation: When to stop extrapolating?* Science & Justice, 56(6):482–491, 2016, ISSN 1355-0306. https://www.sciencedirect.com/science/article/pii/S1355030616300363. xi, 29, 32

[11] Morrison, Geoffrey Stewart: *Tutorial on logistic-regression calibration and fusion:converting a score to a likelihood ratio*. Australian Journal of Forensic Sciences, 45(2):173–197, 2013. https://doi.org/10.1080/00450618.2012.733025. xii, 45, 48, 58

[12] Morrison, Geoffrey Stewart e Norman Poh: *Avoiding overstating the strength of forensic evidence: Shrunk likelihood ratios/bayes factors*. Science & Justice, 58(3):200–218, 2018, ISSN 1355-0306. https://www.sciencedirect.com/science/article/pii/S1355030617301582. xii, 29, 45, 46

[13] Morrison, Geoffrey Stewart, Ewald Enzinger, Vincent Hughes, Michael Jessen, Didier Meuwly, Cedric Neumann, S. Planting, William C. Thompson, David van der Vloed, Rolf J.F. Ypma, Cuiling Zhang, A. Anonymous e B. Anonymous: *Consensus on validation of forensic voice comparison*. Science & Justice, 61(3):299–309, 2021, ISSN 1355-0306. https://www.sciencedirect.com/science/article/pii/S1355030621000083. xii, 51, 52, 53, 54, 63

[14] Ashby, Matthew P. J.: *The Value of CCTV Surveillance Cameras as an Investigative Tool: An Empirical Analysis*. European Journal on Criminal Policy and Research, 23(3):441–459, setembro 2017, ISSN 1572-9869. https://doi.org/10.1007/s10610-017-9341-6. 1

[15] FISWG: *Facial comparison overview and methodology guidelines*, 2019. https://fiswg.org/fiswg_facial_comparison_overview_and_methodology_guidelines_V1.0_20191025.pdf, acesso em 2022-02-12. 1

[16] ENFSI: *Enfsi-bpm-di-01 - best practice manual for facial image comparison*, janeiro 2018. https://enfsi.eu/wp-content/uploads/2017/06/ENFSI-BPM-DI-01.pdf, acesso em 2022-03-05. 1, 2

[17] Willis, S., Andrew Ligertwood, J.J. Molina, Charles Berger, Grzegorz Zadora, Anders Nordgaard, Birgitta Rasmusson, L. Lunt, Christophe Champod, A. Biedermann, Tacha Hicks, Franco Taroni e Xiaochen Zhu: *Enfsi guideline for evaluative reporting in forensic science*, março 2015. https://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf, acesso em 2022-02-12. 2, 8

[18] Forensic Science Australia New Zealand, National Institute of: *An introductory guide to evaluative reporting*, junho 2017. https://www.anzpaa.org.au/

ArticleDocuments/220/An%20Introductory%20Guide%20to%20Evaluative%20Reporting.PDF.aspx, acesso em 2022-02-12. 2

[19] Charles Rodrigues Valente: *Inferência Lógica na Criminalística*. Relatório Técnico, Instituto Nacional de Criminalística, 2018. 2

[20] Domingos Tocchetto e Alberi Espindula: *Criminalística - Procedimentos e Metodologias*. Editora Millennium, Campinas, SP, 5ª edição, 2022, ISBN 978-85-7625-379-2. 3

[21] Norell, Kristin, Klas Brorsson Läthén, Peter Bergström, Allyson Rice, Vaidehi Natu e Alice O'Toole: *The effect of image quality and forensic expertise in facial image comparisons*. Journal of Forensic Sciences, 60(2):331–340, 2015. https://onlinelibrary.wiley.com/doi/abs/10.1111/1556-4029.12660. 3

[22] White, David, P. Jonathon Phillips, Carina Hahn, Matthew Hill e Alice O'Toole: *Perceptual expertise in forensic facial image comparison*. Proceedings. Biological sciences / The Royal Society, 282, setembro 2015. 3

[23] Zoete, J. C. de: *Combining forensic evidence*. 2016, ISBN 9789402803877. https://dare.uva.nl/search?identifier=338e0a8a-176d-4ab7-a370-c022eb4a374c, acesso em 2022-02-12. 3

[24] Schroff, Florian, Dmitry Kalenichenko e James Philbin: *Facenet: A unified embedding for face recognition and clustering*. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun 2015. http://dx.doi.org/10.1109/CVPR.2015.7298682. 3, 15, 16, 34, 48

[25] Wang, Mei e Weihong Deng: *Deep face recognition: A survey*. CoRR, abs/1804.06655, 2018. http://arxiv.org/abs/1804.06655. 3

[26] Grother, Patrick, Mei Ngan e Kayee Hanaoka: *Face Recognition Vendor Test (FRVT) part 2 :: identification*. Relatório Técnico NIST IR 8271, National Institute of Standards and Technology, Gaithersburg, MD, setembro 2019. https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8271.pdf, acesso em 2022-02-12. 3

[27] He, Kaiming, Xiangyu Zhang, Shaoqing Ren e Jian Sun: *Deep residual learning for image recognition*. Em *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 770–778, 2016. 3, 16

[28] Cook, R., I.W. Evett, G. Jackson, P.J. Jones e J.A. Lambert: *A model for case assessment and interpretation*. Science & Justice, 38(3):151–156, 1998, ISSN 1355-0306. https://www.sciencedirect.com/science/article/pii/S1355030698720994. 6, 8

[29] Cook, R., I.W. Evett, G. Jackson, P.J. Jones e J.A. Lambert: *A hierarchy of propositions: deciding which level to address in casework*. Science & Justice, 38(4):231–239, 1998, ISSN 1355-0306. https://www.sciencedirect.com/science/article/pii/S1355030698721173. 6, 8

[30] Mölder, Anna Leida, Isabelle Enlund Åström e Elisabet Leitet: *Development of a score-to-likelihood ratio model for facial recognition using authentic criminalistic data*. Em *2020 8th International Workshop on Biometrics and Forensics (IWBF)*, páginas 1–6, 2020. 9, 20, 25, 48

[31] Meuwly, Didier, Daniel Ramos e Rudolf Haraksim: *A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation*. Forensic Science International, 276:142–153, 2017, ISSN 0379-0738. `https://www.sciencedirect.com/science/article/pii/S0379073816301359`. 9, 50

[32] Morrison, Geoffrey Stewart e Ewald Enzinger: *Score based procedures for the calculation of forensic likelihood ratios – scores should take account of both similarity and typicality*. Science & Justice, 58(1):47–58, 2018, ISSN 1355-0306. `https://www.sciencedirect.com/science/article/pii/S1355030617300849`. 10, 64

[33] Bledsoe, W. W.: *The model method in facial recognition*. Technical report PRI 15, Panoramic Research, Inc., 1964. 11

[34] Bledsoe, W. W. e H. Chan: *A man-machine facial recognition system-some preliminary results*. Technical report PRI 19A, Panoramic Research, Inc., 1965. 11

[35] Bledsoe, W. W.: *Man-machine facial recognition: Report on a large-scale experiment*. Technical report PRI 22, Panoramic Research, Inc., 1966. 11

[36] Bledsoe, W. W.: *Semiautomatic facial recognition*. Technical report SRI Project 6693, Stanford Research Institute, 1968. 11

[37] Kanade, Takeo: *Picture processing system by computer complex and recognition of human faces*, November 1973. 11

[38] Sirovich, L. e M. Kirby: *Low-dimensional procedure for the characterization of human faces*. J. Opt. Soc. Am. A, 4(3):519–524, Mar 1987. `http://josaa.osa.org/abstract.cfm?URI=josaa-4-3-519`. 12

[39] Turk, Matthew e Alex Pentland: *Eigenfaces for recognition*. J. Cognitive Neuroscience, 3(1):71–86, janeiro 1991, ISSN 0898-929X. `https://doi.org/10.1162/jocn.1991.3.1.71`. 12, 20

[40] Etemad, Kamran e Rama Chellappa: *Discriminant analysis for recognition of human face images*. J. Opt. Soc. Am. A, 14(8):1724–1733, Aug 1997. `http://josaa.osa.org/abstract.cfm?URI=josaa-14-8-1724`. 13

[41] Fisher, R.: *The use of multiple measurements in taxonomic problems*. Annals of Human Genetics, 7:179–188, 1936. 13

[42] Liu, Chengjun e H. Wechsler: *Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition*. IEEE Transactions on Image Processing, 11(4):467–476, 2002. 14

[43] Ahonen, T., A. Hadid e M. Pietikainen: *Face description with local binary patterns: Application to face recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(12):2037–2041, 2006. 14

[44] Cao, Zhimin, Qi Yin, Xiaoou Tang e Jian Sun: *Face recognition with learning-based descriptor*. Em *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, páginas 2707–2714, 2010. 14

[45] Lei, Zhen, Matti Pietikäinen e Stan Z. Li: *Learning discriminant face descriptor*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(2):289–302, 2014. 14

[46] Chan, Tsung Han, Kui Jia, Shenghua Gao, Jiwen Lu, Zinan Zeng e Yi Ma: *Pcanet: A simple deep learning baseline for image classification?* IEEE Transactions on Image Processing, 24(12):5017–5032, Dec 2015, ISSN 1941-0042. `http://dx.doi.org/10.1109/TIP.2015.2475625`. 14

[47] Sun, Yi, Yuheng Chen, Xiaogang Wang e Xiaoou Tang: *Deep learning face representation by joint identification-verification*. Em *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, página 1988–1996, Cambridge, MA, USA, 2014. MIT Press. 14, 15

[48] Huang, Gary B., Manu Ramesh, Tamara Berg e Erik Learned-Miller: *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*. Relatório Técnico 07-49, University of Massachusetts, Amherst, October 2007. 14, 15, 17

[49] White, David, Kristin Norell, P. Jonathon Phillips e Alice O'Toole: *Human Factors in Forensic Face Identification*, páginas 195–218. fevereiro 2017. 14

[50] Krizhevsky, Alex, Ilya Sutskever e Geoffrey E Hinton: *Imagenet classification with deep convolutional neural networks*. Em Pereira, F., C. J. C. Burges, L. Bottou e K. Q. Weinberger (editores): *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. `https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf`. 15

[51] Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke e Andrew Rabinovich: *Going deeper with convolutions*. Em *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 1–9, 2015. 16

[52] Zeiler, Matthew D. e Rob Fergus: *Visualizing and understanding convolutional networks*. Em Fleet, David, Tomas Pajdla, Bernt Schiele e Tinne Tuytelaars (editores): *Computer Vision – ECCV 2014*, páginas 818–833, Cham, 2014. Springer International Publishing, ISBN 978-3-319-10590-1. 16

[53] Kemelmacher-Shlizerman, Ira, Steven M. Seitz, Daniel Miller e Evan Brossard: *The megaface benchmark: 1 million faces for recognition at scale*. Em *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 16, 17

[54] Deng, Jiankang, Jia Guo, Niannan Xue e Stefanos Zafeiriou: *Arcface: Additive angular margin loss for deep face recognition.* Em *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 16, 35, 48

[55] Yang, Shuo, Ping Luo, Chen Change Loy e Xiaoou Tang: *Wider face: A face detection benchmark.* Em *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 16

[56] Deng, Jiankang, Jia Guo, Evangelos Ververas, Irene Kotsia e Stefanos Zafeiriou: *Retinaface: Single-shot multi-level face localisation in the wild.* Em *CVPR*, 2020. 16

[57] Howard, Andrew G., Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, M. Andreetto e Hartwig Adam: *Mobilenets: Efficient convolutional neural networks for mobile vision applications.* ArXiv, abs/1704.04861, 2017. 16

[58] Sandler, M., Andrew G. Howard, Menglong Zhu, A. Zhmoginov e Liang Chieh Chen: *Mobilenetv2: Inverted residuals and linear bottlenecks.* 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, páginas 4510–4520, 2018. 16

[59] Zhang, X., X. Zhou, Mengxiao Lin e Jian Sun: *Shufflenet: An extremely efficient convolutional neural network for mobile devices.* 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, páginas 6848–6856, 2018. 16

[60] Tan, Mingxing e Quoc V. Le: *Efficientnet: Rethinking model scaling for convolutional neural networks.* ArXiv, abs/1905.11946, 2019. 16

[61] Yi, Dong, Zhen Lei, Shengcai Liao e Stan Z. Li: *Learning face representation from scratch.* CoRR, abs/1411.7923, 2014. `http://arxiv.org/abs/1411.7923`. 17

[62] Parkhi, Omkar M., Andrea Vedaldi e Andrew Zisserman: *Deep face recognition.* Em *Proceedings of the British Machine Vision Conference (BMVC)*, páginas 41.1–41.12. BMVA Press, September 2015, ISBN 1-901725-53-7. `https://dx.doi.org/10.5244/C.29.41`. 17

[63] Guo, Yandong, Lei Zhang, Yuxiao Hu, Xiaodong He e Jianfeng Gao: *Ms-celeb-1m: A dataset and benchmark for large-scale face recognition.* CoRR, abs/1607.08221, 2016. `http://arxiv.org/abs/1607.08221`. 17

[64] An, Xiang, Xuhan Zhu, Yang Xiao, Lan Wu, Ming Zhang, Yuan Gao, Bin Qin, Debing Zhang e Ying Fu: *Partial FC: training 10 million identities on a single machine.* CoRR, abs/2010.05222, 2020. `https://arxiv.org/abs/2010.05222`. 17

[65] Oliveira Jr, L. L. de: *Captura e alinhamento de imagens: Um banco de faces brasileiro.* Technical report, Centro Universitário da FEI, 2006. 17

[66] Neves, Joao, Juan Moreno e Hugo Proença: *Quis-campi: an annotated multi-biometrics data feed from surveillance scenarios.* IET Biometrics, 7, outubro 2017. 17

[67] Zeinstra, Chris G., Raymond N.J. Veldhuis, Luuk J. Spreeuwers, Arnout C.C. Ruifrok e Didier Meuwly: *Forenface: a unique annotated forensic facial image dataset and toolset*. IET Biometrics, 6(6):487–494, 2017. https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-bmt.2016.0160. 17

[68] Eidinger, Eran, Roee Enbar e Tal Hassner: *Age and gender estimation of unfiltered faces*. IEEE Transactions on Information Forensics and Security, 9(12):2170–2179, 2014. 17, 38

[69] Robinson, Joseph P, Gennady Livitz, Yann Henon, Can Qin, Yun Fu e Samson Timoner: *Face recognition: Too bias, or not too bias?* Em *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, páginas 1–10, 2020. 17

[70] Chen, Jun Cheng, Rajeev Ranjan, Swami Sankaranarayanan, Amit Kumar, Ching Hui Chen, Vishal M. Patel, Carlos D. Castillo e Rama Chellappa: *Unconstrained Still/Video-Based Face Verification with Deep Convolutional Neural Networks*. International Journal of Computer Vision, 126(2-4):272–291, abril 2018, ISSN 0920-5691, 1573-1405. http://link.springer.com/10.1007/s11263-017-1029-3, acesso em 2022-03-28. 20

[71] Ding, Changxing e Dacheng Tao: *Trunk-Branch Ensemble Convolutional Neural Networks for Video-Based Face Recognition*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(4):1002–1014, abril 2018, ISSN 0162-8828, 2160-9292, 1939-3539. https://ieeexplore.ieee.org/document/7917252/, acesso em 2022-03-28. 20

[72] Gong, Sixue, Yichun Shi e Anil Jain: *Low Quality Video Face Recognition: Multi-Mode Aggregation Recurrent Network (MARN)*. Em *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, páginas 1027–1035, Seoul, Korea (South), outubro 2019. IEEE, ISBN 978-1-72815-023-9. https://ieeexplore.ieee.org/document/9022380/, acesso em 2022-03-28. 20

[73] Ranjan, Rajeev, Ankan Bansal, Hongyu Xu, Swami Sankaranarayanan, Jun Cheng Chen, Carlos D. Castillo e Rama Chellappa: *Crystal loss and quality pooling for unconstrained face verification and recognition*, 2019. 20

[74] Yang, Jiaolong, Peiran Ren, Dongqing Zhang, Dong Chen, Fang Wen, Hongdong Li e Gang Hua: *Neural aggregation network for video face recognition*. Em *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 5216–5225, 2017. 20, 64

[75] Liu, Yu, Junjie Yan e Wanli Ouyang: *Quality aware network for set to set recognition*. Em *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 4694–4703, 2017. 20

[76] Gong, Sixue, Yichun Shi e Anil K. Jain: *Video face recognition: Component-wise feature aggregation network (c-fan)*, 2019. 20

[77] Foreman, Lindsey A., Christophe Champod, Ian W. Evett, James A. Lambert e Susan Pope: *Interpreting dna evidence: A review.* International Statistical Review, 71:473–495, 2007. 20

[78] Meuwly, Didier e Andrzej Drygajlo: *Forensic speaker recognition based on a bayesian framework and gaussian mixture modelling (gmm).* Em *Odyssey*, 2001. 20

[79] Zeinstra, C, Didier Meuwly, Arnout Ruifrok, Raymond Veldhuis e Luuk Spreeuwers: *Forensic face recognition as a means to determine strength of evidence: A survey.* Forensic science review, 30:21–32, janeiro 2018. 20

[80] Rodriguez, Andrea Macarulla, Zeno Geradts e Marcel Worring: *Calibration of score based likelihood ratio estimation in automated forensic facial image comparison.* Forensic Science International, 334:111239, 2022, ISSN 0379-0738. https://www.sciencedirect.com/science/article/pii/S037907382200069X. 21

[81] Tippett, C.F., V.J. Emerson, M.J. Fereday, F. Lawton, A. Richardson, L.T. Jones e Miss S.M. Lampert: *The evidential value of the comparison of paint flakes from sources other than vehicles.* Journal of the Forensic Science Society, 8(2):61–65, 1968, ISSN 0015-7368. https://www.sciencedirect.com/science/article/pii/S0015736868704424. 22, 53

[82] Mandasari, Miranti Indar, Manuel Günther, Roy Wallace, Rahim Saeidi, Sébastien Marcel e David A. van Leeuwen: *Score calibration in face recognition.* IET Biometrics, 3(4):246–256, 2014. https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-bmt.2013.0066. 23, 59

[83] Vogt, Robbie e Sridha Sridharan: *Explicit modelling of session variability for speaker verification.* Computer Speech & Language, 22(1):17–38, 2008, ISSN 0885-2308. https://www.sciencedirect.com/science/article/pii/S0885230807000277. 23

[84] Brümmer, Niko e Johan du Preez: *Application-independent evaluation of speaker detection.* Computer Speech & Language, 20(2):230–275, 2006, ISSN 0885-2308. https://www.sciencedirect.com/science/article/pii/S0885230805000483, Odyssey 2004: The speaker and Language Recognition Workshop. 23, 51, 52

[85] Ali, Tauseef: *Biometric Score Calibration for Forensic Face Recognition.* Tese de Doutoramento, University of Twente, junho 2014, ISBN 978-90-365-3689-9. 24, 25, 34, 48

[86] Ramos, Daniel, Joaquin Gonzalez-Rodriguez, Grzegorz Zadora e Colin Aitken: *Information-theoretical assessment of the performance of likelihood ratio computation methods.* Journal of Forensic Sciences, 58(6):1503–1518, 2013. https://onlinelibrary.wiley.com/doi/abs/10.1111/1556-4029.12233. 25, 29

[87] Ruifrok, A.C.C., P. Vergeer e Andrea Macarulla Rodrigues: *From facial images of different quality to score based lr.* Forensic Science International, 332:111201, 2022, ISSN 0379-0738. https://www.sciencedirect.com/science/article/pii/S0379073822000317. 28, 29

[88] Brummer, Niko: *Measuring, refining and calibrating speaker and language information extracted from speech*. Tese de Doutoramento, Stellenbosch University, December 2010. 30, 31

[89] Serengil, Sefik Ilkin e Alper Ozpinar: *Lightface: A hybrid deep face recognition framework*. Em *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, páginas 23–27. IEEE, 2020. https://doi.org/10.1109/ASYU50717.2020.9259802. 35

[90] Zhang, Kaipeng, Zhanpeng Zhang, Zhifeng Li e Yu Qiao: *Joint face detection and alignment using multitask cascaded convolutional networks*. IEEE Signal Processing Letters, 23(10):1499–1503, 2016. 35

[91] Guo, Jia, Jiankang Deng, Alexandros Lattas e Stefanos Zafeiriou: *Sample and computation redistribution for efficient face detection*. Em *International Conference on Learning Representations*, 2022. https://openreview.net/forum?id=RhB1AdoFfGE. 35

[92] Terhörst, Philipp, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner e Arjan Kuijper: *Ser-fiq: Unsupervised estimation of face image quality based on stochastic embedding robustness*. Em *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, páginas 5650–5659, 2020. 39, 49

[93] Ruifrok, A.C.C., P. Vergeer e Andrea Macarulla Rodrigues: *From facial images of different quality to score based lr*. Forensic Science International, página 111201, 2022, ISSN 0379-0738. https://www.sciencedirect.com/science/article/pii/S0379073822000317. 39, 49

[94] Jin, Chi, Ruochun Jin, Kai Chen e Yong Dou: *A community detection approach to cleaning extremely large face database*. Computational Intelligence and Neuroscience, 2018:1–10, 2018. https://doi.org/10.1155/2018/4512473. 40

[95] Silverman, B. W.: *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, London, 1986. 44

[96] Buch-larsen, Tine, Jens Perch Nielsen, Montserrat Guillén e Catalina Bolancé: *Kernel density estimation for heavy-tailed distributions using the champernowne transformation*. Statistics, 39(6):503–516, 2005. https://doi.org/10.1080/02331880500439782. 44

[97] Ramos-Castro, Daniel, Joaquin Gonzalez-Rodriguez e Javier Ortega-Garcia: *Likelihood ratio calibration in a transparent and testable forensic speaker recognition framework*. Em *2006 IEEE Odyssey - The Speaker and Language Recognition Workshop*, páginas 1–8, 2006. 44, 51

[98] Gonzalez-Rodriguez, Joaquin, Phil Rose, Daniel Ramos, Doroteo T. Toledano e Javier Ortega-Garcia: *Emulating dna: Rigorous quantification of evidential weight in transparent and testable forensic speaker recognition*. IEEE Transactions on Audio, Speech, and Language Processing, 15(7):2104–2115, 2007. 44

[99] Morrison, Geoffrey Stewart: *Likelihood-ratio forensic voice comparison using parametric representations of the formant trajectories of diphthongs*. The Journal of the Acoustical Society of America, 125(4):2387–2397, 2009. https://doi.org/10.1121/1.3081384. 44

[100] Morrison, Geoffrey Stewart: *A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: Multivariate kernel density (mvkd) versus gaussian mixture model–universal background model (gmm–ubm)*. Speech Communication, 53(2):242–256, 2011, ISSN 0167-6393. https://www.sciencedirect.com/science/article/pii/S016763931000155X. 44

[101] Mansournia, Mohammad Ali, Angelika Geroldinger, Sander Greenland e Georg Heinze: *Separation in Logistic Regression: Causes, Consequences, and Control*. American Journal of Epidemiology, 187(4):864–870, agosto 2017, ISSN 0002-9262. https://doi.org/10.1093/aje/kwx299. 45

[102] King, DE: *Dlib-ml: A machine learning toolkit in journal of machine learning research vol. 10*. 2009. 48

[103] Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot e E. Duchesnay: *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research, 12:2825–2830, 2011. 48

[104] Meuwly, Didier: *Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique*. Tese de Doutoramento, Université de Lausanne, Faculté de droit et des sciences criminelles, 2000. 53

[105] Kim, Minchul, Feng Liu, Anil Jain e Xiaoming Liu: *Cluster and aggregate: Face recognition with large probe set*. Em *Advances in Neural Information Processing Systems*, 2022. 64