UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

GUSTAVO ALEXANDRE TUCHLINOWICZ NUNES

APLICAÇÃO DA TÉCNICA OMNI UTILIZANDO MATCHING COMO MEDIDA DE DISTÂNCIA PARA REDUÇÃO DO ESPAÇO DE BUSCA EM BANCOS DE DADOS DE BIOMETRIA DIGITAL INFANTIL

PATO BRANCO

2024

GUSTAVO ALEXANDRE TUCHLINOWICZ NUNES

APLICAÇÃO DA TÉCNICA OMNI UTILIZANDO MATCHING COMO MEDIDA DE DISTÂNCIA PARA REDUÇÃO DO ESPAÇO DE BUSCA EM BANCOS DE DADOS DE BIOMETRIA DIGITAL INFANTIL

Application of the OMNI Technique Using Matching as a Distance Measure to Reduce the Search Space in Infant Digital Biometrics Databases

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Computação do Curso de Bacharelado em Engenharia de Computação da Universidade Tecnológica Federal do Paraná.

Orientador: Prof. Dr. Marcelo Teixeira

Coorientador: Prof. Dr. Ives Rene Venturini Pola

PATO BRANCO

2024

GUSTAVO ALEXANDRE TUCHLINOWICZ NUNES

APLICAÇÃO DA TÉCNICA OMNI UTILIZANDO MATCHING COMO MEDIDA DE DISTÂNCIA PARA REDUÇÃO DO ESPAÇO DE BUSCA EM BANCOS DE DADOS DE BIOMETRIA DIGITAL INFANTIL

Trabalho de Conclusão de Curso de Graduação apresentado como requisito para obtenção do título de Bacharel em Engenharia de Computação do Curso de Bacharelado em Engenharia de Computação da Universidade Tecnológica Federal do Paraná.

Data de aprovação: 20/junho/2024

Prof. Dr. Marcelo Teixeira
Universidade Tecnológica Federal do Paraná, Campus Pato Branco

---

Prof. Dr. Ives Rene Venturini Pola
Universidade Tecnológica Federal do Paraná, Campus Pato Branco

---

Prof. Dr. Luis Carlos Ferreira Bueno
Universidade Tecnológica Federal do Paraná, Campus Pato Branco

---

Prof. Ms. Geri Natalino Dutra
Universidade Tecnológica Federal do Paraná, Campus Pato Branco

---

Profa. Dra. Viviane Dal Molin
Universidade Tecnológica Federal do Paraná, Campus Pato Branco

PATO BRANCO

2024

A todos que acreditaram em mim, obrigado por
todo o suporte e compreensão.

AGRADECIMENTOS

Agradeço também aos meus orientadores por todo o suporte prestado durante o desenvolvimento deste trabalho.

À minha família, agradeço por toda a paciência nos momentos em que reclamei, e também por todo o suporte nos momentos de dificuldade.

À minha namorada Luana, só você sabe quantas noites passei em claro desenvolvendo este trabalho. Agradeço pela companhia e por todo o apoio que me deu. Sou eternamente grato pelo privilégio de tê-la como minha companheira.

E aos amigos que a universidade me deu, agradeço por cada brincadeira, cada lista de cálculo e cada jantar do grupo. Nunca será um adeus, e sim um até logo. Neste momento tão especial, lhes digo: eu ganhei.

Agradece-se à InfantID por seu inestimável apoio operacional ao grupo de pesquisa em biometria neonatal na UTFPR. Sua parceria tem sido fundamental para o avanço de exploração de tecnologias biométricas no atendimento neonatal. O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior -Brasil (CAPES) - Código de Financiamento 001 e pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ).

RESUMO

A biometria digital tem se mostrado eficaz na identificação de recém-nascidos, ajudando a evitar trocas de bebês em hospitais e a facilitar o controle de vacinas. No entanto, à medida que os bancos de dados de digitais infantis aumentam, surgem problemas de lentidão nas buscas durante a identificação. Geralmente, a busca é realizada de forma exaustiva, comparando cada elemento da base de dados um a um. Em bases de dados muito grandes, isso pode se tornar um empecilho significativo. Portanto, é necessário estudar métodos para acelerar a recuperação dessas digitais, sendo este o foco do presente trabalho. Existem vários estudos sobre melhorias de busca em larga escala; entretanto, até o momento, nenhum foi testado em bases de biometria digital infantil. Assim, o objetivo deste trabalho é aplicar a técnica OMNI, que apresenta uma redução de espaço de busca sub-linear em relação ao tamanho da base de dados. Foi utilizado o *matching* de digitais como medida de distância, em especial o algoritmo MCC, algo ainda não testado para biometria digital infantil. Os experimentos realizados apresentaram reduções de até 92% no espaço de busca, mantendo uma taxa de acerto de 100%. Além disso, experimentalmente foi possível utilizar o MCC como medida de distância em um espaço métrico. Para trabalhos futuros, recomenda-se testar diferentes algoritmos de *matching* e adicionar mais digitais à base de dados para verificar a robustez e a eficiência da abordagem nesses casos.

Palavras-chave: banco de dados; biometria; recém-nascidos; recuperação de dados; impressão digital.

ABSTRACT

Fingerprint biometrics have proven effective in identifying newborns, helping to prevent baby swaps in hospitals and facilitating vaccine tracking. However, as databases of infant fingerprints grow, issues of search slowness during identification arise. Typically, the search is performed exhaustively, comparing each element in the database one by one. In very large databases, this can become a significant hindrance. Therefore, it is necessary to study methods to speed up the retrieval of these fingerprints, which is the focus of the present work. There are several studies on large-scale search improvements; however, to date, none have been tested on infant fingerprint databases. Thus, the objective of this work is to apply the OMNI technique, which shows a sub-linear reduction in search space relative to the database size. Fingerprint matching was used as a distance measure, specifically the MCC algorithm, something not yet tested for infant fingerprint biometrics. The experiments conducted showed reductions of up to 92% in search space, maintaining a 100% accuracy rate. Additionally, it was experimentally possible to use MCC as a distance measure in a metric space. For future work, it is recommended to test different matching algorithms and add more fingerprints to the database to verify the robustness and efficiency of the approach in these cases.

Keywords: database; biometrics; newborns; data retrieval; fingerprint.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

Siglas

GPU	Unidade de Processamento Gráfico, do inglês *Graphics Processing Unit*
KNN	K-Vizinhos Mais Próximos, do inglês *K-Nearest Neighbors*
MCC	Código Cilíndrico de Minúcias, do inglês *Minutia Cylinder-Code*
FAR	Taxa de Falsas Aceitações, do inglês *False Acceptance Rate*
FRR	Taxa de Falsas Rejeições, do inglês *False Rejection Rate*
TAR	Taxa de Aceitações Verdadeiras, do inglês *True Acceptance Rate*
TRR	Taxa de Rejeições Verdadeiras, do inglês *True Rejection Rate*
TPIR	Taxa de Positivos Identificados como Negativos, do inglês *True Positives Identified as Negatives Rate*
CMC	Característica Cumulativa de Correspondência, do inglês *Cumulative Match Characteristic*
SQL	Linguagem de Consulta Estruturada, do inglês *Structured Query Language*
NOSQL	Não apenas SQL, do inglês *Not only SQL*
XML	Linguagem de Marcação Extensível, do inglês *eXtensible Markup Language*
JSON	Notação de Objeto JavaScript, do inglês *JavaScript Object Notation*
SQL	Linguagem de Consulta Estruturada, do inglês *Structured Query Language*
TCC	Trabalho de Conclusão de Curso

SUMÁRIO

1 INTRODUÇÃO

De acordo com um estudo realizado pela Organização das Nações Unidas (2022), de 2020 em diante, 35% das vítimas de tráfico humano são crianças. Nesse contexto, dispor de métodos eficazes e ágeis para a identificação de indivíduos é fundamental. Isso impacta desde os aspectos de proteção individual, até a detecção precoce de crianças recém-nascidas em situações vulneráveis, controle vacinal, combate à fome, etc. Para lidar com a complexidade desses controles e sua amplitude geográfica, pesquisadores têm concentrado esforços na área do reconhecimento biométrico, envolvendo características como a análise facial, a analise de impressões digitais e o reconhecimento da íris, entre outras (Preciozzi *et al.*, 2020).

Em se tratando de crianças recém-nascidas, a identificação biométrica se torna particularmente importante e tecnicamente mais desafiadora em comparação com a de adultos. Isso ocorre porque a coleta da impressão biométrica costuma ser dificultada pelas características dos dedos das crianças, que podem conter elementos como óleo e umidade, além de apresentarem um desenho biométrico ainda em formação, muitas vezes com poucas ou nenhuma característica registrável por meio dos *scanners* usuais de coleta. Além disso, o tamanho da impressão digital é reduzido, tornando a manipulação adequada do dedo mais difícil, entre outros empecilhos (Engelsma *et al.*, 2021).

Nesse sentido, o processamento computacional desponta como um diferencial para mitigar parte das limitações com a coleta da estampa biométrica de crianças. Técnicas como a de super-resolução, por exemplo, são capazes de receber como entrada uma imagem limitada em termos de identificação e devolver como saída imagens com melhores características (Wang; Chen; Hoi, 2021). Outra oportunidade de melhoria no processo de identificação emerge da extração, armazenamento e resgate apropriado de características associadas à estampa biométrica (Primo *et al.*, 2019). Este trabalho é particularmente focado em melhorias ao processo de armazenamento eletrônico e nos métodos de extração de informações a partir de bases de dados construídas para aplicações de biometria.

Ao implementar um sistema de identificação biométrica, é comum necessitar de uma ampla busca na base de dados a fim de que se encontre um *match*, ou seja, uma digital que coincida com aquela a ser validada. Nesse processo de busca, cada entrada (imagem da digital a ser validada) é processada e uma pesquisa percorre todo o banco de dados buscando o *match*, caracterizando o que é conhecido como busca 1 para N. Uma busca 1 para N é um método do tipo exaustivo e, como tal, requer um dispêndio computacional considerável, fazendo com que o tempo de resposta da consulta seja inversamente proporcional à quantidade de dados presentes na base. Como, no caso da identificação biométrica, as bases costumam ter volumes substanciais de dados, encontrar o *match* pode ser inviável frente ao tempo de identificação de que se dispõe, em geral na casa de segundos. Portanto, ao mesmo tempo em que é essencial dispor de bancos de dados biométricos, o processo de busca por validação costuma ser um entrave para sua viabilidade e usabilidade prática.

Dentre as possibilidades para acelerar o processo de busca temos por exemplo técnicas baseadas em árvores de decisão, como a *Random Forest*, que utiliza uma estrutura hierárquica para dividir o problema em subproblemas menores, porém, ela pode ser sensível a dados com grande volume dimensional (Ghosh; Cabrera, 2022). Também existe a possibilidade de utilização de GPUs (Unidades de Processamento Gráfico), entretanto, uma desvantagem notável é a necessidade de hardware específico e caro. Nesse sentido, uma alternativa promissora emerge da busca por similaridade em bases de dados complexos.

A busca por similaridade foca em encontrar um conjunto de possíveis soluções, e não apenas a solução exata. Essa abordagem é utilizada quando a correspondência exata pode ser difícil devido a variações nos dados, ruído ou imprecisões. Um exemplo de técnica com busca por similaridade é a OMNI, apresentada por Traina *et al.* (2001). Essa técnica consiste em delimitar regiões de busca dentro das quais haja certo grau de similaridade entre certas variáveis a serem consideradas no problema. Como efeito, a técnica OMNI possibilita estabilizar o tempo de busca independente do tamanho do conjunto de dados, perdendo assim parte da complexidade da busca 1 para N, sob o esforço adicional, mas em geral viável, de computar as regiões de busca.

Dada a natureza do problema de *matching* 1 para N e o perfil da busca OMNI, seria esperado que ambos pudessem ser integrados para a redução da dimensionalidade da busca. Essa conjectura ainda não foi explorada na literatura e sustenta o objetivo principal desse trabalho.

O objetivo geral deste trabalho é validar a eficácia da técnica OMNI em um banco de dados dedicado à biometria digital infantil, que inclui características da estampa biométrica que possam facilitar o reconhecimento e aperfeiçoar o processo de busca, e também verificar a utilização do *matching* como medida de distância em um espaço métrico. Além disso, ao concentrar-se especialmente na amplificação do desempenho e aceleração do processo de correspondência biométrica, a abordagem proposta envolve a recuperação por similaridade, utilizando a técnica OMNI para que o referido banco possa ser utilizado para coleta contínua de digitais, a fim de tornar a abordagem aplicável em larga escala. Para validar a eficácia do método proposto, serão realizados testes utilizando bases de dados reais de biometria infantil. Os testes serão feitos aplicando a busca sequencial antes e depois de aplicar a OMNI.

1.1 Objetivos específicos

Os seguintes objetivos específicos são propostos para o cumprimento do objetivo geral:

- Implementação do banco de dados
- Implementação do *matching* no banco de dados
- Implementação da função OMNI no banco de dados

- Validação da utilização do *matching* como medida de distância
- Comparação do método proposto com a busca sequencial
- Validação da eficiência do método proposto

O presente trabalho possui um total de seis capítulos, incluindo a introdução. O capítulo 2 trata dos trabalhos relacionados, apresentando uma abordagem sobre outros trabalhos que apresentam similaridade e embasamento para o atual trabalho. O capítulo 3 contém o referencial teórico, onde serão apresentados os conceitos necessários para o entendimento do trabalho. Em seguida será apresentado o capítulo 4, o qual explica como foi desenvolvida a parte prática do trabalho, incluindo os materiais e os métodos e um fluxograma do trabalho. No capítulo 5 serão apresentados os resultados obtidos e em seguida no capítulo 6 a conclusão do trabalho.

2 TRABALHOS RELACIONADOS

Este campo de pesquisa ganha destaque à medida que a necessidade de identificação e proteção das crianças se torna uma prioridade crescente em muitas sociedades. A detecção precisa e eficaz de impressões digitais infantis é um componente fundamental no auxílio a questões de segurança e bem-estar infantil, bem como em situações de emergência, como sequestros e desaparecimentos.

No contexto da velocidade de busca, diferentes abordagens foram feitas para tentar melhorar esse processo, como, por exemplo, a utilização de GPUs (Cappelli; Ferrara; Maltoni, 2015) e até mesmo o acesso sequencial (Indrawan; Sitohang; Akbar, 2012). Dentro da área de inteligência artificial existem os algoritmos de busca, como busca em profundidade, busca em largura, busca gulosa, busca A*, entre outros. Esses algoritmos visam uma melhoria na velocidade em que uma busca é realizada em uma determinada problemática, e alguns deles foram utilizados na literatura com este propósito, como Iloanusi e Osuagwu (2011) que utilizou o algoritmo *K-Means* para digitais e também e Mehrotra *et al.* (2010) que utilizou a variação *K-Means Fuzzy* para reconhecimento de assinaturas. Outra alternativa estudada é o algoritmo de redes siamesas, essa abordagem tem se destacado especialmente na área de identificação biométrica, como no reconhecimento de faces. A ideia central por trás das redes siamesas é treinar duas redes idênticas simultaneamente, utilizando pares de dados de entrada. Durante o treinamento, o objetivo é fazer com que essas redes gerem representações similares para itens semelhantes e representações distintas para itens diferentes, dessa maneira esse algoritmo tem a capacidade de aprender características discriminativas, tornando-as eficazes na comparação de semelhanças entre diferentes instâncias. As redes siamesas demonstraram bons resultados para identificação de rostos comparado a outros métodos, e, como citado por Wu *et al.* (2017), o esperado é que quanto maior a base de dados de treinamento, melhores os resultados.

No sentido de buscas por similaridades em larga escala, já foram estudadas algumas técnicas, onde a maioria é baseada na indexação das imagens e dos vetores métricos, ou baseadas na divisão do espaço de busca em regiões menores.

O artigo de Mhatre *et al.* (2005), que utilizou uma base de dados de assinaturas e imagens de mãos, faz uma abordagem baseada em árvores em que a base de dados é separada em compartimentos, sendo que cada compartimento possui dados similares, dessa forma, no caso da busca 1:N, no lugar de pesquisar exaustivamente todo o banco de dados, o método proposto faz com que a busca inicialmente seja feita procurando o compartimento correspondente do dado, diminuindo assim o tamanho do espaço de busca. Ao utilizar a geometria das mãos e as assinaturas, aplicando o método proposto foi possível reduzir o espaço de busca para apenas 5% do tamanho original.

Também utilizando árvores, Chen, Bouman e Allebach (1997) fizeram um teste de busca explorando a técnica de Quantização Vetorial Estruturada em Árvore (TSVQ) e a incorporando a outras duas técnicas, a busca por ramificação e limite, e a desigualdade triangular. O obje-

tivo do conjunto era acelerar a pesquisa em grandes bancos de dados de imagens. O método proposto reduz a computação necessária para localizar imagens que melhor correspondem a uma imagem de consulta fornecida pelo usuário. Para realização dos experimentos foi utilizada uma base de dados contendo 10.000 imagens, foi criado um vetor de característica de tamanho 211 formado a partir de histogramas de cor, textura e informações de borda. É concluído que o método proposto pode acelerar significativamente a busca em estruturas TSVQ, porém uma colocação importante é que mesmo a busca exata sendo possível, resultados melhores são alcançados ao procurar apenas por imagens similares.

No âmbito de biometria digital, Lumini, Maio e Maltoni (1997) abordam o problema da recuperação de impressões digitais latentes em um grande banco de dados. A justificativa principal para o desenvolvimento do referido artigo é que as abordagens tradicionais adotam a classificação exclusiva de impressões digitais, e ao passar para o contexto de busca em larga escala, essa técnica não se mostrava efetiva. Para tentar resolver o problema, foi proposta uma técnica de classificação contínua, em que o princípio básico é a caracterização das impressões digitais com vetores em um espaço multidimensional. Os resultados obtidos mostram que um desempenho melhor pode ser alcançado, em ambas as metodologias, por meio da abordagem contínua. Os autores ainda destacam que embora a classificação contínua não permita realizar algumas tarefas, se classificarmos as impressões digitais apenas para melhorar a eficiência da recuperação, a abordagem contínua é uma melhor opção quando comparada à abordagem exclusiva.

Já Boer, Bazen e Gerez (2001) exploram a eficácia de três técnicas diferentes de indexação baseadas em múltiplas características: a estimativa do campo direcional registrado, o *FingerCode* e as tríades de pontos de minúcia. As três técnicas mostraram melhoras em relação à busca exaustiva, sendo que o método do campo direcional apresentou melhores resultados. Um dos motivos para isso é que o cálculo do *FingerCode* leva cerca de 40 segundos no processador utilizado, enquanto o cálculo do campo direcional leva apenas 15 segundos. O método baseado nas tríades apresentou resultados inferiores aos outros dois. Ao final do artigo é realizado um teste onde é criado um novo método, que é baseado na combinação dos três métodos anteriores. Esse novo método apresentou resultados ainda melhores que os individuais, permitindo que os bancos de dados tenham um tamanho até 100 vezes maiores, mantendo a mesma Taxa de Falsas Aceitações (FAR) e a Taxa de Falsas Rejeições (FRR).

A variedade de métodos que focam na celeridade da busca em larga escala vai além destes citados, um exemplo é a técnica OMNI, por Traina *et al.* (2001). A técnica OMNI é baseada no conceito de selecionar focos da base de dados que irão servir como referência para os cálculos de distância. O artigo apresenta como encontrar os melhores focos e como utilizá-los, e também destaca que os focos aumentam a poda de cálculos de distância durante o processamento de consulta. Para realizar os testes da técnica apresentada, foram utilizadas três bases de dados diferentes:

- Palavras em inglês: Uma base de dados com 25.143 objetos da língua inglesa, para essa base de dados a função de distância escolhida foi da *Ledit*.
- Faces: 11.900 vetores de faces, em que cada vetor tinha 16 dimensões. Foi utilizada a função de distância Euclidiana.
- Sintéticos: Uma base de dados sintética com 250.000 objetos, e nesse caso também foi utilizada a distância Euclidiana.

O estudo apresentou resultado satisfatório para a técnica, tendo uma melhoria de até 10 vezes na velocidade de busca, e também é destacado o fato de que a técnica apresentou uma boa escalabilidade, tendo um comportamento sub-linear com o aumento da base de dados.

Apesar da variedade de métodos que visam melhorar a velocidade da busca em larga escala, incluindo o método OMNI discutido neste capítulo, é importante observar que, até o momento, não foram realizados estudos que testaram a eficácia da técnica OMNI em bancos de dados de impressões digitais, tanto de adultos quanto de crianças. Dada a crescente demanda por métodos eficientes de busca em larga escala em bancos de dados de impressões digitais infantis e a capacidade demonstrada pelo método OMNI de escalabilidade para bancos de dados maiores e eficiência na redução do tempo de busca, este trabalho visa aplicar o método em uma base de dados de impressões digitais infantis para avaliar sua aplicabilidade e eficácia.

O quadro 1, apresentado a seguir, mostra um resumo dos trabalhos relacionados a buscas por similaridades em larga escala.

Quadro 1 – Trabalhos relacionados

Autor	Título	Contexto	Abordagem
MHATRE, A. J. et al.	Efficient search and retrieval in biometric databases	Assinaturas e imagens de mãos	Árvores
CHEN, J.-Y.; BOUMAN, C.; ALLEBACH, J.	Fast image database search using tree-structured	Imagens em geral	Quantização Vetorial Estruturada em Árvore
LUMINI, A.; MAIO, D.; MALTONI, D.	Continuous versus exclusive classification for fingerprint retrieval	Impressões digitais latentes	Classificação contínua
BOER, J. D.; BAZEN, A. M.; GEREZ, S. H	Indexing fingerprint databases based on multiple features	Impressões digitais de adultos	Campo direcional registrado, o FingerCode e as tríades de pontos de minúcia
TRAINA, A. et al.	Similarity search without tears: the OMNI-family of all-purpose access methods	Palavras, faces e dados sintéticos	OMNI

Fonte: Autoria própria (2024).

3 REFERENCIAL TEÓRICO

O presente capítulo apresenta os conceitos necessários para o entendimento do trabalho e também para resolver o problema tratado no capítulo anterior.

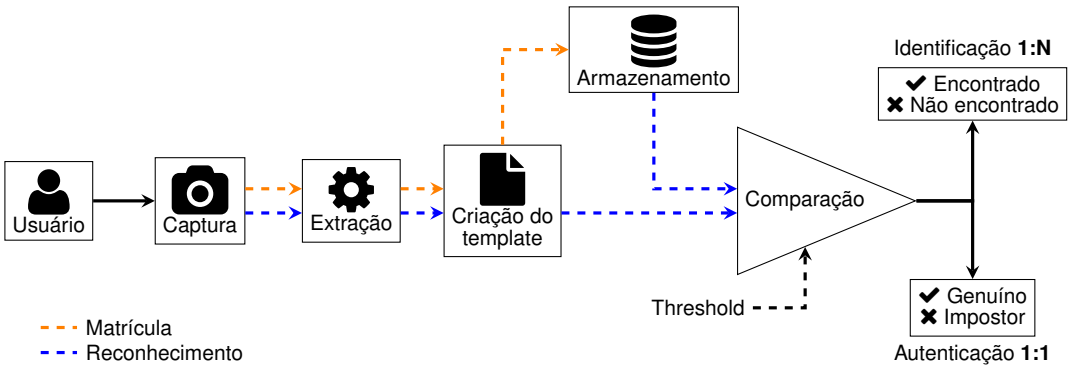
3.1 Biometria

Podemos definir biometria como a utilização de características próprias de um indivíduo para proceder à sua autenticação e/ou identificação (Magalhães, 2003). Existem vários tipos de biometria, por exemplo íris, digital, facial e até mesmo voz. Cada uma tem suas particularidades e pode ser usada para diferentes objetivos. De forma geral a biometria é utilizada para identificação de pessoas, como por exemplo em celulares ou em urnas eletrônicas. O presente trabalho tem foco na biometria digital, que se mostrou como uma ótima escolha para identificação de recém-nascidos visto que evidências biológicas sugerem que os padrões de impressões digitais estão fisiologicamente presentes nos dedos humanos desde o nascimento, além disso, as impressões digitais são a biometria mais conveniente, aceitável e economicamente viável para ser capturada de bebês (Engelsma *et al.*, 2019). A biometria digital é baseada em características especificas que as digitais possuem, os padrões de vales, cristas e minúcias tornam cada digital única. O reconhecimento por meio de biometria digital é feito com base nas características citadas anteriormente, e para conseguir estudar essas características é necessário primeiramente extraí-las, e para isso é utilizado um sistema de extração.

3.2 Sistema de reconhecimento

Como apresentado na Figura 1, um sistema de reconhecimento pode ser resumido no passo a passo de captura, extração, armazenamento, comparação, identificação e autenticação. Primeiramente são coletadas amostras biométricas dos indivíduos, essas amostras são processadas e armazenadas como modelos. Em seguida é realizado o processo de extração, no qual são extraídas as características das amostras. A partir do momento em que um indivíduo deseja acessar algum recurso seguro, ele deve inserir sua digital em um mecanismo de captura, essa etapa é denominada verificação. Após a verificação existem duas etapas que podem ser realizadas, a autenticação, que é o processo de verificar a correspondência dessa digital com os modelos armazenados de um indivíduo específico (1:1), e a outra etapa possível é a identificação, na qual essa digital é comparada a um banco de dados contendo modelos de vários indivíduos (1:N) (Southier *et al.*, 2023).

Figura 1 – Esquema generalizado de um sistema de reconhecimento



Fonte: Adaptado de Southier *et al.* (2023).

3.3 Coleta

Durante o processo de coleta, um sensor é empregado para quantificar as características biométricas, gerando assim uma representação da referida informação. No caso da impressão digital, é produzida uma imagem bidimensional que retrata o dedo. No caso desses dados específicos, a qualidade da imagem é imediatamente afetada pela resolução, pela taxa de quadros e pela sensibilidade do sensor óptico. O sensor óptico desempenha um papel crucial no sistema de identificação, pois é ele que inicia todo o processo. De forma geral é necessária uma resolução de 500 pontos por polegada, ou 500 ppi, na maior parte das aplicações biométricas, para que se tenha um processamento e uma comparação bem-sucedidos (Jain; Ross; Nandakumar, 2011).

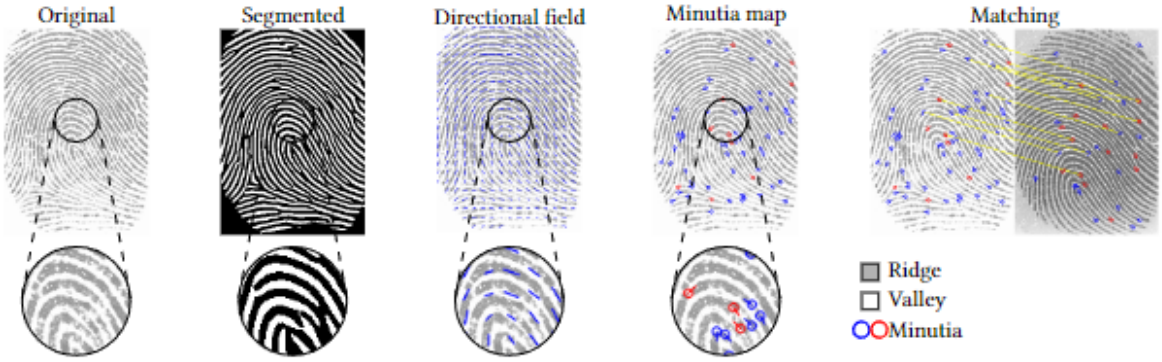
3.4 Extratores de características

Para separar as características das imagens originais são utilizados extratores de características, esses extratores são métodos computacionais que possuem a imagem original como parâmetro e devolvem as características escolhidas. Dentre os extratores podemos citar o campo direcional, bitmap, segmentação e mapa de minúcias.
A Figura 2 apresenta exemplos de características que podem ser extraídas de uma digital, e também apresenta uma imagem simplificada do processo de matching, que será tratado em um dos capítulos seguintes.

3.4.1 Segmentação

De acordo com Albanez Marcos Aurélio Batista (2016), segmentação de imagem é o processo de agrupar pixels ou conjuntos de pixels que tenham uma mesma propriedade. Uma outra definição é que de forma generalizada, segmentação consiste em subdividir uma imagem de entrada em partes constituintes da mesma.

Figura 2 – Exemplo das características extraídas da digital



Fonte: Southier *et al.* (2023).

Os autores destacam que existem várias técnicas de segmentação, mas não existe um método que funcione para todos os tipos as imagens. No âmbito de impressões digitais a segmentação é tida como uma tarefa relativamente simples, uma vez que o fundo não contém ruídos. Nesse sentido, a segmentação é a divisão da imagem em uma área de "primeiro plano" da impressão digital e uma área de "segundo plano" de ruído (Bazen; Gerez, 2002). Portanto, algumas medidas como a variância local do nível de cinza e o nível médio local de cinza podem ser usadas (Ratha; Chen; Jain, 1995).

3.4.2 Campo direcional

O campo direcional é o conjunto das orientações locais de uma digital, e a orientação local da crista no ponto (j,i) é o ângulo $\theta_{j,i} \in [0, 180°[$ que as cristas da impressão digital formam com o eixo horizontal em uma vizinhança arbitrariamente pequena centrada em (j,i).

Para cada pixel, estima-se a orientação local a partir do gradiente [Gx,Gy] , que já foi calculado na etapa de segmentação (Bazen; Gerez, 2002).

A orientação da crista é estimada como ortogonal à orientação do gradiente, sendo a média em uma janela W .

$$G_{xx} = \sum_{W} G_x^2 \tag{1}$$

$$G_{yy} = \sum_{W} G_y^2 \tag{2}$$

$$G_{xy} = \sum_{W} G_x G_y \tag{3}$$

$$\theta = \frac{\pi}{2} + \frac{\text{phase}(G_{xx} - G_{yy} 2G_{xy})}{2} \tag{4}$$

Para cada orientação, também calcula-se um valor de confiança (*strength*), que mede o quanto todos os gradientes em W compartilham a mesma orientação.

$$\text{strength} = \frac{\sqrt{(G_{xx} - G_{yy})^2 + (2G_{xy})^2}}{G_{xx} + G_{yy}} \tag{5}$$

3.4.3 Mapa de minúcias

Um mapa de minúcias é uma representação visual ou estrutural das características distintas encontradas em uma impressão digital. Para identificar uma impressão digital como única é preciso mapear as minúcias, que são formas presentes nas digitais. Diversos tipos de minúcias estão presentes, e geralmente podemos identificar entre 40 a 100 dessas características em uma imagem de alta qualidade da impressão digital (Natosafe, 2022). Minúcias são pontos de interesse únicos na superfície das impressões digitais, e elas são classificadas em dois tipos principais: bifurcações e terminações, mas também existem outros tipos de minúcias como as ilhotas e encerros.

- Bifurcações: São pontos em que a linha da impressão digital se divide em duas partes.
- Ponta de linha/Terminações: São pontos em que uma linha da impressão digital termina abruptamente.

Para criar um mapa de minúcias, um algoritmo de processamento de imagens digitaliza a impressão digital e, em seguida, identifica e marca esses pontos de bifurcação e terminação (Espinosa-Duro, 2002). O resultado é uma representação gráfica ou matricial que descreve a disposição e a localização dessas minúcias na impressão digital. Um mapa de minúcias pode ser visto na Figura 2 apresentada anteriormente.

A Figura 3 mostra visualmente os diferentes tipos de minúcias. Esses conceitos são importantes pois o mapa de minúcias é basicamente o conjunto dessas minúcias encontradas em uma digital, e que por sua vez será utilizado para o mapeamento das estruturas locais.

3.4.4 Estrutura local

O conceito de estrutura local é baseado na análise do conjunto de minúcias em uma região específica de uma impressão digital, a Figura 4 apresenta uma representação gráfica de uma estrutura local. No contexto do método de *matching* clássico, simplificadamente, essa

Figura 3 – Exemplificação de minúcias em uma digital

Fonte: Natosafe (2022).

estrutura local consiste no conjunto de minúcias em um cilindro, cuja base e altura estão relacionadas à informação espacial e direcional, respectivamente, criado ao redor de uma minúcia particular (Cappelli; Ferrara; Maltoni, 2010).

Com as estruturas locais mapeadas, pode-se realizar o *matching* utilizando-as.

3.4.5 Matching

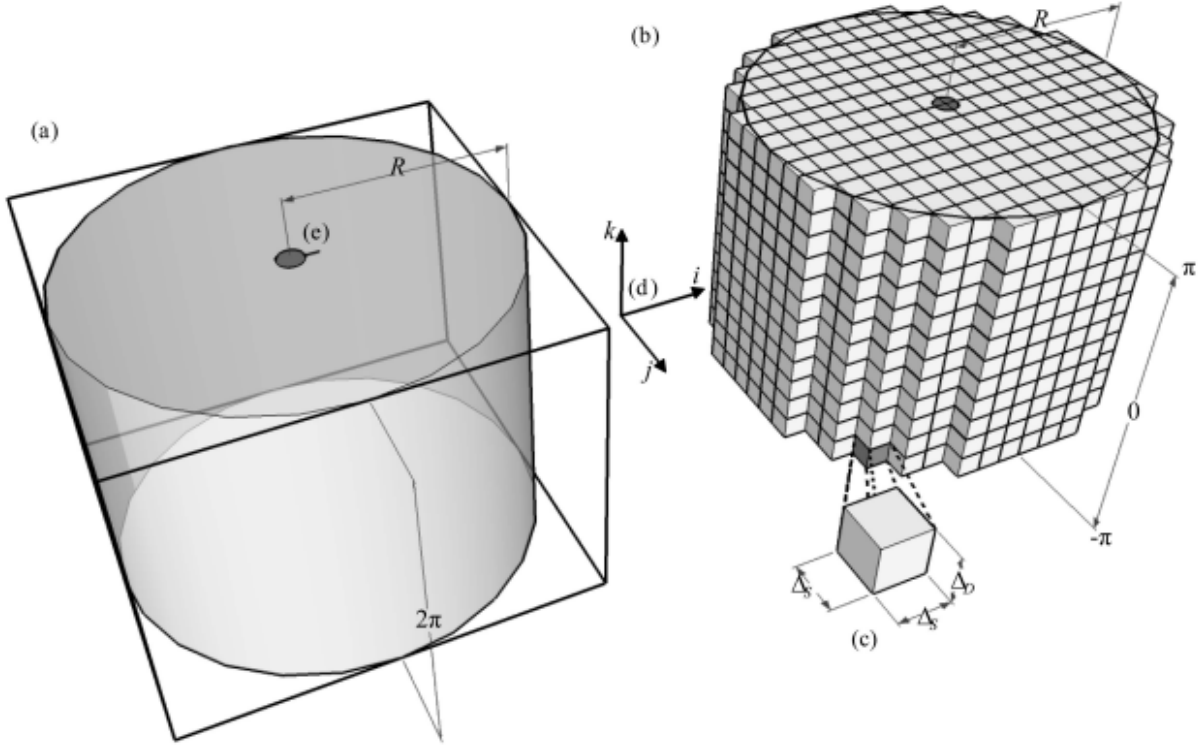
Matching é o conceito de verificar a correspondência de duas informações. No âmbito da biometria digital, existem vários métodos para *matching*, sendo o principal método utilizado atualmente o *matching* clássico, baseado no conceito de *Minutia Cylinder-Code* (MCC). Este método faz uso das estruturas locais das digitais (Cappelli; Ferrara; Maltoni, 2010).

De maneira simplificada, o método seleciona uma minúcia e sua respectiva estrutura local e, em seguida, percorre as minúcias de uma segunda digital, buscando uma que tenha uma estrutura local semelhante à primeira. Esse procedimento é repetido até que seja satisfeita uma condição de aceitação ou que o número de verificações seja excedido.

A aceitação desse método é determinada pela inclinação das linhas que ligam as minúcias das duas digitais. Se uma porcentagem das linhas for paralela, significa que as digitais são correspondentes; se as linhas tiverem inclinações diferentes, então isso indica que as duas digitais pertencem a pessoas diferentes ou que não é possível chegar a uma conclusão.

A precisão de um método de *matching* de forma geral é determinada pelas taxas FRR, Taxa de Aceitações Verdadeiras (TAR), FAR e Taxa de Rejeições Verdadeiras (TRR), essas

Figura 4 – Representação gráfica de uma estrutura local



Fonte: Cappelli, Ferrara e Maltoni (2010).

quatro taxas estão interligadas e dependem do limite (*threshold*) selecionado (Southier *et al.*, 2023).

- FRR: A porcentagem de usuários genuínos que foram rejeitados.
- TAR: O complemento da FRR (1-FRR), a porcentagem de usuários genuínos aceitos.
- FAR: Parcela de impostores aceitos.
- TRR: Parcela de impostores rejeitados, sendo TRR igual a 1 - FAR.

Em resumo, o TAR mede a capacidade de reconhecimento de correspondências válidas, enquanto o FAR mede a capacidade de evitar falsas aceitações. Ambos são parâmetros essenciais para avaliar o desempenho de sistemas de *matching* em biometria digital.

Já para a identificação, cada digital é comparada com as outras disponíveis no banco de dados/galeria, e então os resultados são organizados e classificados em ordem. Em seguida, é medida a Taxa de Identificação Verdadeira Positiva (TPIR), que determina a posição em que ocorre uma correspondência verdadeira. A taxa TPIR calcula a probabilidade de identificação correta dentre as primeiras K posições, e para mostrar os resultados da identificação é utilizada a Característica Cumulativa de Correspondência (CMC) (Southier *et al.*, 2023).

Destaca-se que a coleta pode ser feita mais de uma vez e para um ou mais dedos, além de que cada um dos conceitos apresentados acima pode ser armazenado em um banco de

dados para cada uma das imagens da coleta, para que futuros estudos não precisem extraí-los novamente. Nota-se que com o passar dos anos os bancos de dados de digitais podem acabar se tornando muito grandes, e o armazenamento em pastas pode não ser a maneira mais eficiente de guardar estes dados para estudos.

3.5 Banco de dados

Banco de dados são conjuntos de dados que são relacionados entre si, podendo ser registros de pessoas, locais, datas, entre outros. De forma generalizada um banco de dados é uma maneira de guardar informações sobre algo, por exemplo podemos guardar dados de uma biblioteca, em que iriamos salvar os nomes dos livros, autores, quem emprestou o livro, horário do empréstimo e qualquer outra informação relevante para o gerenciamento eficiente da biblioteca. Desse modo podemos ter um registro temporal de todos os empréstimos feitos na biblioteca. De maneira semelhante podemos guardar os dados das coletas de digitais, guardando dados como local de coleta, horário da coleta, imagem coletada e também as características dos extratores citados anteriormente. Exitem muitas variáveis na hora de se escolher um banco de dados, este pode ser um banco convencional ou "não apenas SQL"(NoSQL), pode ser relacional ou não relacional, e cada um tem suas vantagens e desvantagens.

3.5.1 Relacional

Um banco de dados relacional, como o nome sugere, é um sistema de gerenciamento de dados em relações, no qual uma relação é um termo matemático para tabela. Cada tabela é um conjunto de linhas, em que cada linha possui o mesmo conjunto de colunas nomeadas, e cada coluna é um tipo de dado específico (PostgreSQL, 2023). No contexto de coletas de digitais, um banco de dados relacional pode armazenar informações como informações biométricas associadas a dados pessoais.

3.5.2 Modelos de Bancos de Dados

Bancos de dados podem ser classificados em diferentes modelos, atendendo a necessidades específicas de organização e flexibilidade. Aqui estão alguns modelos comuns:

- **Banco Estruturado:** Organiza dados conforme um esquema rígido e definido. Útil para dados biométricos consistentes em formato predefinido.
- **Banco Não Estruturado:** Caracterizado pela ausência de formato de dados predefinido. Ideal para dados biométricos com variações de formato e resolução.

- **Banco Semiestruturado:** Combina estrutura e flexibilidade, permitindo a organização parcialmente definida, como em documentos Linguagem de Marcação Extensível (XML) ou Notação de Objeto JavaScript (JSON). Vantajoso para informações biométricas com alguma estrutura e variações de detalhes ou metadados.

3.6 Busca 1 para N

A busca 1 para N ou também busca em grande escala é o tipo de busca realizado em grandes bancos de dados, popularmente utiliza-se o método de busca exaustiva ou busca por força bruta. Esse tipo de busca é útil para banco pequenos, nos quais não há tanta necessidade de velocidade, entretanto, quando o tamanho do banco se torna muito grande, a busca exaustiva pode ser muito ineficiente (Boer; Bazen; Gerez, 2001). Desse modo faz-se necessária a criação de novas formas de busca, e a que será testada nesse trabalho é a redução do espaço de busca utilizando a técnica OMNI. Para isso é necessário primeiro entender os conceitos de espaço métrico e funções de distância.

3.7 Espaço métrico

Uma métrica em um domínio S é uma função d : S × S $\rightarrow$ R+ que associa a cada par ordenado de elementos (s1, s2) um número real d(s1,s2) chamado de distância de s1 a s2, e que atenda as propriedades definidas no espaço métrico (Lima, 2020).

Um espaço métrico, denotado como M, é uma estrutura matemática composta por um par ordenado (S, d), em que:

- S é um conjunto que contém elementos ou objetos.
- d é uma métrica, também conhecida como função de distância, que associa pares de elementos do conjunto S a valores reais não negativos, refletindo a dissimilaridade entre esses elementos.

3.8 Função de Distância

Para avaliar a similaridade entre dois elementos em um determinado domínio, é essencial utilizar uma função de distância. Esta função recebe como parâmetros um par de elementos do conjunto e retorna um valor que reflete a dissimilaridade entre eles. Quanto mais próximo este valor estiver de zero, mais semelhantes são considerados os elementos em questão.

Para que uma distância seja válida em um espaço métrico, ela precisa respeitar as seguintes propriedades:

1. **Não-negatividade**:

$$d(x, y) \geq 0 \quad \text{para todos } x, y \in X \tag{6}$$

2. **Identidade do indiscernível**:

$$d(x, y) = 0 \quad \text{se e somente se } x = y \tag{7}$$

3. **Simetria**:

$$d(x, y) = d(y, x) \quad \text{para todos } x, y \in X \tag{8}$$

4. **Desigualdade triangular**:

$$d(x, z) \leq d(x, y) + d(y, z) \quad \text{para todos } x, y, z \in X \tag{9}$$

Dentre essas propriedades, destaca-se a propriedade de desigualdade triangular, em que segundo Pola (2010), essa propriedade é fundamental para evitar cálculos desnecessários, auxiliando na 'poda' das subárvores, e é particularmente relevante para a técnica OMNI. A utilização de funções de distância e a observância da desigualdade triangular contribuem para a eficiência do processo de indexação e recuperação de dados em espaços métricos complexos, como os abordados por essa técnica. Funções de distância comuns incluem distância euclidiana, distância de Manhattan, distância de Jaccard, entre outras.

3.9 Técnica OMNI

A Técnica OMNI é uma abordagem avançada de indexação de banco de dados complexos, projetada para otimizar as consultas baseadas em funções de distância. Ela se concentra em organizar os dados de forma a permitir recuperações eficientes em espaços métricos, especialmente quando a proximidade entre objetos é essencial. De maneira simplificada, a técnica OMNI seleciona focos distintos da base de dados, em que cada foco é um elemento que se destaca da base de dados, normalmente selecionado como sendo o elemento mais distante (de acordo com a função de distância) de todos os outros elementos.

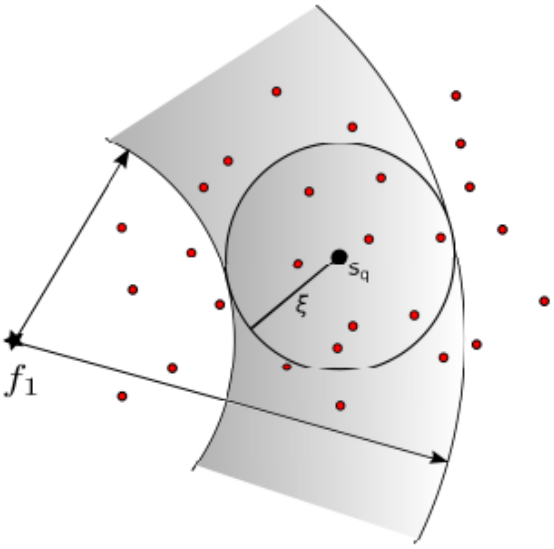
De acordo com Traina *et al.* (2006), uma base de focos OMNI pode ser definida como:

Definição 1. *Seja um espaço métrico $M = (S, d)$, uma base de focos OMNI é um conjunto $F = \{f_1, f_2, \ldots, f_l \mid f_k \in S, f_k \neq f_j, l \leq N\}$, em que cada f_k é um foco (ou ponto focal) de S, l é o número de focos da base de focos e N é o número de elementos da base de dados.*

Dada uma determinada entrada, a técnica OMNI mede a distância entre essa entrada e os focos, e assim consegue gerar uma sub-região da base de dados na qual a entrada tende

a estar. Dessa forma o espaço de busca é reduzido significativamente. Na Figura 5, a área sombreada representa o espaço de busca selecionado pela técnica OMNI quando utilizado apenas um foco, em que f1 é o foco, Sq é o elemento central da base de dados e ξ é o raio de abrangência.

Figura 5 – Consulta por abrangência pela técnica OMNI utilizando um foco



.

Fonte: Matsui (2018).

Nota-se na imagem que os elementos que ficam na região sombreada são aqueles em que sua distância para o foco f1 fica entre determinados valores, valores estes que são definidos como a distância de Sq para o foco, mais ou menos o raio escolhido para a busca. Após a aplicação da técnica OMNI, tem-se um espaço de busca reduzido, no qual a digital correspondente está inclusa, sendo assim, pode-se utilizar algum algoritmo para encontrar a mesma.

4 MATERIAIS E MÉTODOS

O presente capítulo descreve os materiais e métodos usados no trabalho.

4.1 Materiais

Dentre os materiais que foram utilizados para desenvolver este trabalho estão o PostgreSQL, a linguagem de consulta SQL, a linguagem de programação Python e a base de dados biométricos infantis.

4.1.1 PostgreSQL

O PostgreSQL é um sistema de banco de dados objeto-relacional poderoso de código aberto com mais de 35 anos de desenvolvimento ativo, o que lhe conferiu uma sólida reputação em termos de confiabilidade, robustez de recursos e desempenho (POSTGRESQL, 2023). Para esse trabalho foi utilizada a versão 16 do PostgreSQL.

4.1.2 SQL

A Linguagem de Consulta Estruturada (SQL) é uma linguagem de programação interativa que permite aos usuários encontrar e modificar informações em bancos de dados (Mohn, 2023). A linguagem SQL é essencial para a interação com o banco de dados, permitindo a criação, atualização, consulta e recuperação de informações de forma eficiente. Neste trabalho, o SQL foi utilizado para formular consultas que extraem dados específicos da base de dados, facilitando a análise e interpretação dos resultados, e também foi utilizada para implementar as funções necessárias durante o desenvolvimento.

4.1.3 Python

Para implementação do MCC dentro do banco de dados, foi utilizada a linguagem de programação Python na versão 3.11 (compatível com a versão 16 do PostgreSQL). Em conjunto ao Python, foi utilizada a biblioteca Numpy, sua utilização se faz necessária para realizar cálculos rápidos com vetores, permitindo melhor desempenho do algoritmo de *matching*.

4.1.4 Base de imagens

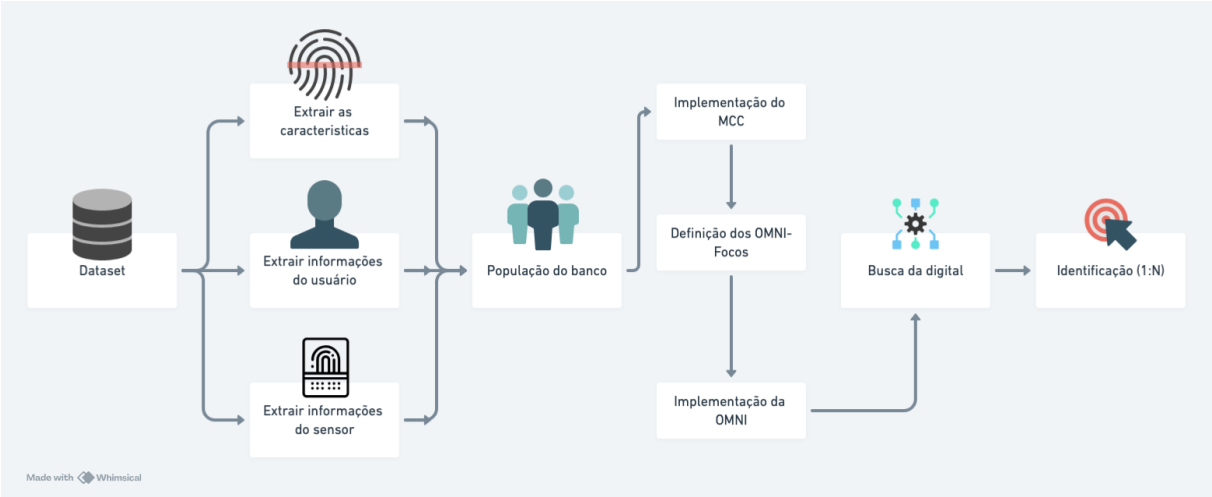
O desenvolvimento deste trabalho visa a criação de um banco que seja temporal e que possa ser incrementado em futuras pesquisas. Para o desenvolvimento do trabalho, foi utilizada

a base de dados do CMDB, citada por Basak *et al.* (2017), que possui cerca de 5000 imagens. Foram removidas todas as imagens que tinham menos que 20 minúcias após a extração do mapa de minúcias, visto que essas digitais apresentavam erros no resultado do *matching*, o que poderia comprometer o teste da técnica OMNI, totalizando assim 2.758 imagens. No futuro esse mesmo banco será atualizado recorrentemente com as imagens coletadas pelo grupo de pesquisa em biometria neonatal [1].

4.2 Métodos

O trabalho pode ser dividido em seis etapas principais, sendo elas a criação do banco de dados, extração das características da base de dados, implementação do MCC, definição dos focos, implementação da função OMNI e a busca da digital no banco de dados. A Figura 6 apresenta um fluxograma do desenvolvimento dos métodos do presente trabalho. Estes métodos serão comentados nas próximas seções.

Figura 6 – Fluxo dos métodos



Fonte: Autoria própria (2024).

4.3 Criação do banco de dados

Nessa etapa foram escolhidas quais informações seriam armazenadas no banco de dados e também foi criado o banco no software PostgreSQL. Foram selecionadas três categorias principais de informações, sendo elas: "Informações da pessoa", "Informações da coleta"e "Informações da imagem".

[1] Estudo aprovado com Certificado de Apresentação de Apreciação Ética 73791023.7.0000.0177 na Plataforma Brasil

Para cada categoria foram escolhidas as informações especificas que seriam armazenadas, nas quais cada informação seria um atributo da respectiva categoria.

4.3.1 Informações da pessoa

Armazenar as informações da pessoa que se está coletando a digital é muito importante principalmente para a questão de temporalidade do banco de dados, em que em um cenário ideal uma mesma pessoa teria várias digitais no banco, em momentos diferentes, sendo assim, essas informações são valiosas para comparações de tempos em tempos.

4.3.2 Informações da coleta

Na parte de informações da coleta estão as entidades de coleta inicial e re-coleta, nas quais são guardadas informações como o horário da coleta, peso e altura no momento da coleta, semanas de gestação e o local da coleta. Além disso, é a partir da coleta que conectamos o banco à sua maior seção, as informações da imagem.

4.3.3 Informações da imagem

Essa é a maior seção do banco, em que são armazenadas as imagens e extrações de características da mesma. Nessa seção, além das imagens e extrações, também se tem as entidades de minúcias, sensor e dedos, e também as duas entidades principais da técnica OMNI, sendo elas a entidade de minúcias e a dos pivôs.

Para a criação do banco foi utilizada a linguagem SQL no software PostgreSQL, e tanto o código quanto uma imagem do modelo do banco estão presentes no capítulo de resultados.

4.4 Implementação do MCC

A característica escolhida para calcular a dissimilaridade entre as digitais foi o mapa de minúcias, entretanto, a maioria das distâncias só funcionam em determinadas condições, condições estas que geralmente incluem que os elementos estejam ordenados igualmente e tenham o mesmo tamanho, o que nem sempre acontece com o mapa de minúcias das digitais. Uma mesma digital pode ter duas coletas em que o número de minúcias varia, sendo assim não seria possível utilizar uma distância clássica para calcular a dissimilaridade das digitais. Dessa forma, foi decidido testar um algoritmo de *matching* como medida de distância, neste caso o MCC.

Como explicado anteriormente, o MCC é um algoritmo que compara as digitais baseado nas minúcias e retorna um resultado entre 0 e 1, sendo 1 correspondência completa entre as

digitais, e 0 sendo nenhuma correspondência. Porém, como apresentado no referencial teórico, para ser uma medida de distância válida, o MCC precisa respeitar as 4 propriedades de uma distância métrica. Desta forma nesta etapa, antes de implementar o MCC, foi necessário mostrar que o MCC respeita as 4 propriedades. Para isso foram selecionados 3 mapas de minúcias de digitais escolhidas aleatoriamente no banco de dados e então foram realizados experimentos de verificação das propriedades, e a validação destas 4 propriedades pode ser encontrada no capítulo 5, no qual estão os resultados do trabalho.

Após a validação do MCC como medida de distância, foi então iniciada a implementação do MCC no banco de dados. Primeiramente foi necessário instalar o Python dentro do PostgreSql, permitindo assim utilizar tanto a linguagem quanto suas bibliotecas diretamente dentro do banco de dados. Com o Python instalado, foi instalada em seguida a biblioteca *numpy* e criada a extensão no Postgre para utilizar o Python.

Com o Python e o *numpy* configurados, se iniciou a implementação do MCC no banco de dados. Primeiramente são definidos os parâmetros que serão utilizados pelo MCC para criar as estruturas locais que serão comparadas posteriormente. Os valores escolhidos seguem os valores do artigo original, e são apresentados na tabela a seguir.

Tabela 1 – Valores dos parâmetros

Parâmetro	Descrição	Valor
R	Raio do cilindro (em pixels)	70
Ns	Número de células dentro do cilindro	16
MinNp	Número mínimo de minúcias	4
MaxNp	Número máximo de minúcias	12
Up, Tp	Parâmetros da função sigmoid	20, 2/5

Fonte: Adaptado de Cappelli, Ferrara e Maltoni (2010).

A Tabela 1 apresenta os valores dos parâmetros utilizados no cálculo do MCC. Nessa tabela, R representa o tamanho do raio da estrutura local, N_s indica o número de células internas à estrutura local, $MinNp$ e $MaxNp$ correspondem ao número mínimo e máximo de minúcias utilizadas pelo método, e Up e Tp são utilizados para calcular o número de minúcias a serem empregadas no método. Os valores selecionados seguem as diretrizes do artigo original que descreve o método.

Com estas funções e parâmetros definidos, foi então implementado o código que define as estruturas locais de uma digital, baseado no mapa de minúcias da mesma. Essa etapa é importante pois como explicado anteriormente, o MCC utiliza as estruturas locais da digital para realizar o cálculo da distância. Com as estruturas locais calculadas, se inicia o processo do *matching* em si. Nessa etapa são calculadas as distâncias euclidianas normalizadas em pares entre as estruturas locais da digital de entrada e da digital que está sendo comparada.

Com essas distâncias definidas, é então feita a seleção dos pares de minúcias com as menores distâncias entre si, que serão utilizados para definir o *score*(distância) entre as duas digitais.

Por fim, com os pares definidos, é então calculada a distância em si, finalizando assim o processo de *matching* do MCC. O código do MCC, ou seja, da distância entre duas digitais, foi salvo no banco de dados como uma função para facilitar a utilização na técnica OMNI e a versão completa do código pode ser encontrada no apêndice C. Toda a implementação foi baseada no artigo original do MCC.

4.5 Definição dos OMNI focos

Com a base de dados inserida no banco e o MCC implementado, para selecionar os focos foi utilizado o artigo original da OMNI como referência. No artigo é apresentada uma forma de decidir a quantidade e como encontrar os focos. Para a base de dados em questão, foram escolhidos um total de três focos. Para o primeiro foco, foi selecionada uma digital aleatória da base de dados, e então utilizando o MCC em toda a base restante, comparando a digital aleatória com toda a base de dados. Dessa forma, ao ordenar por distância, foi possível encontrar a digital mais distante, ou seja, menos semelhante à digital que foi escolhida aleatoriamente. Essa digital selecionada é o primeiro foco. Para o segundo foco, foi feito o mesmo procedimento, mas agora utilizando o primeiro foco. Sendo assim, o segundo foco é simplesmente a digital mais distante do primeiro foco. O terceiro foco é escolhido com base nos dois primeiros. O terceiro foco é uma digital que tem uma distância similar tanto para o primeiro quanto para o segundo foco. O cálculo utilizado para encontrar o terceiro foco está presente no artigo original da OMNI (Traina *et al.* (2006)). Com os focos definidos, foi então utilizada a função do MCC para calcular as distâncias de todas as digitais para os 3 focos, e estes valores foram salvos na tabela "Minúcias" no banco de dados, pois são utilizados pela OMNI no momento da poda, que será explicado na sessão seguinte.

4.6 Implementação da técnica OMNI

Com os focos definidos, foi então implementado o código da busca utilizando a técnica OMNI. A OMNI foi inserida em uma função no banco de dados que consiste de 3 etapas:

- Cálculo da distância para os focos
- Poda baseada nas distâncias obtidas
- Ordenação por distância

A função criada recebe como parâmetros uma matriz representando o mapa de minúcias a ser buscado, recebe também o raio de busca e o valor máximo de distância permitido. A partir do mapa de minúcias, é aplicado o MCC para calcular a distância desse mapa até os 3 focos da base de dados, e o resultado desse cálculo é salvo em variáveis temporárias internas da função. Com as distâncias calculadas, inicia-se a poda da base de dados. Esse processo é feito a partir

de um comando de SELECT no banco de dados, no qual passamos as distâncias calculadas no passo anterior na cláusula WHERE, na qual definimos que as distâncias das digitais do banco de dados devem respeitar as seguintes regras:

$$d(qc, foco1) - raio <= dx1 <= d(qc, foco1) + raio \tag{10}$$

$$d(qc, foco2) - raio <= dx2 <= d(qc, foco2) + raio \tag{11}$$

$$d(qc, foco2) - raio <= dx3 <= d(qc, foco2) + raio \tag{12}$$

Em que qc representa a digital de entrada e dx1, dx2 e dx3 representam as distâncias já calculadas e salvas de todas as digitais do banco de dados até os respectivos focos 1, 2 e 3. Ou seja, o que a cláusula WHERE está fazendo é remover todas as digitais em que suas distâncias para os focos diferem da distância da digital buscada para os focos em um certo valor, tanto para mais quanto para menos.

Por exemplo, supondo que se tem 3 digitais no banco, e que suas distâncias para os focos já foram calculadas e estão apresentadas na tabela 2.

Tabela 2 – Distâncias das digitais de exemplo para os focos

Digital	Distância para o foco 1	Distância para o foco 2	Distância para o foco 3
1	0.40	0.35	0.34
2	0.20	0.28	0.30
3	0.25	0.30	0.37

Fonte: Autoria própria (2024).

Caso esteja sendo feita a busca com um raio de 0.05 e a digital de entrada tenha suas distâncias para os focos calculadas sendo 0.21, 0.27 e 0.31, para os focos 1, 2 e 3 respectivamente, então teríamos o seguinte:

$$0.21 - 0.05 <= dx1 <= 0.21 + 0.05 \tag{13}$$

$$0.27 - 0.05 <= dx2 <= 0.27 + 0.05 \tag{14}$$

$$0.31 - 0.05 <= dx3 <= 0.31 + 0.05 \tag{15}$$

Desta forma, nessa busca de exemplo, seria mantida apenas a digital 2, pois é a única das 3 que respeita as 3 cláusulas, sendo assim, nessa etapa a digital 2 representa um falso positivo, ou seja, uma potencial correspondência para a digital que esta sendo buscada. Nesta etapa, todos os falsos positivos da base de dados serão selecionados, e a partir destes são selecionados o identificador e a distância até a digital de entrada. Também é aplicado um filtro para que a distância não seja maior que o parâmetro de valor máximo de distância, sendo assim, qualquer falso positivo que esteja além desse valor será cortado, em outras palavras, digitais que tem pouca semelhança mas estavam entre os falso positivos são removidas da busca.

Essa medida é necessária pois o *matching* ainda apresenta algumas dificuldades para calcular a distância entre digitais infantis, entretanto por não ser o foco desse trabalho, não será discutido aqui, apenas vale ressaltar que futuros estudos que melhorem o *matching* irão eventualmente melhorar essa poda. Para finalizar a função, foi ordenado por distância, colocando as digitais mais semelhantes primeiro, e retornado o resultado da consulta, tendo assim o novo espaço de busca ordenado. Resultado este que já inclui uma busca sequencial, visto que estamos retornando em ordem. O código completo da função OMNI implementada no banco de dados pode ser encontrado no apêndice D.

4.7 Busca da digital no banco de dados

Com as etapas anteriores concluídas, foi iniciada a etapa de busca da digital no banco de dados. Nesta etapa foram realizados testes do método apresentado e os resultados encontrados estão apresentados a seguir. Para realizar os testes, devido às limitações de base de dados disponíveis, foi selecionada uma digital que possuía duas imagens de seções de coleta diferentes que eram do mesmo dedo da mesma pessoa. Dessa forma é possível realizar buscas de uma dessas digitais esperando encontrar a sua correspondência. Os resultados apresentados na seção seguinte mostram o espaço de busca reduzido, o quanto este foi reduzido para cada raio, e também os tempos de busca das digitais.

5 RESULTADOS EXPERIMENTAIS

Esse capítulo apresenta os resultados obtidos neste trabalho, incluindo a etapa de modelagem do banco de dados, a inserção das imagens neste banco, a execução da técnica OMNI sobre essa base, bem como os testes de performance obtidos a partir da técnica OMNI e seu tratamento estatístico.

5.1 Modelagem conceitual do banco de dados

A Figura 7 apresenta o modelo conceitual do banco de dados, que contém as entidades citadas anteriormente, com destaque para a tabela "Minúcias"e a tabela "Pivots", que são as tabelas utilizadas pela técnica OMNI.

Por questões de simplificação de visualização, nessa imagem foram removidos os tamanhos de alguns atributos.

A entidade principal do modelo é a "Coleta Inicial", onde idealmente será armazenada a primeira coleta do recém-nascido. Caso seja possível realizar mais coletas para essa mesma pessoa, as informações entram na tabela de "Recoleta", permitindo assim várias coletas para uma mesma pessoa. Isso é importante para estudos onde o foco é a variação da digital com o passar do tempo.

Na tabela de "Extrações" foram adicionados os atributos "Tipos de Extração" e "Resultado da Extração", desta forma é possível fazer diferentes extrações e todas ficam salvas no mesmo lugar, referentes à mesma digital. Este formato permite que caso novos tipos de extração sejam criados no futuro, o banco ainda possa receber os mesmos.

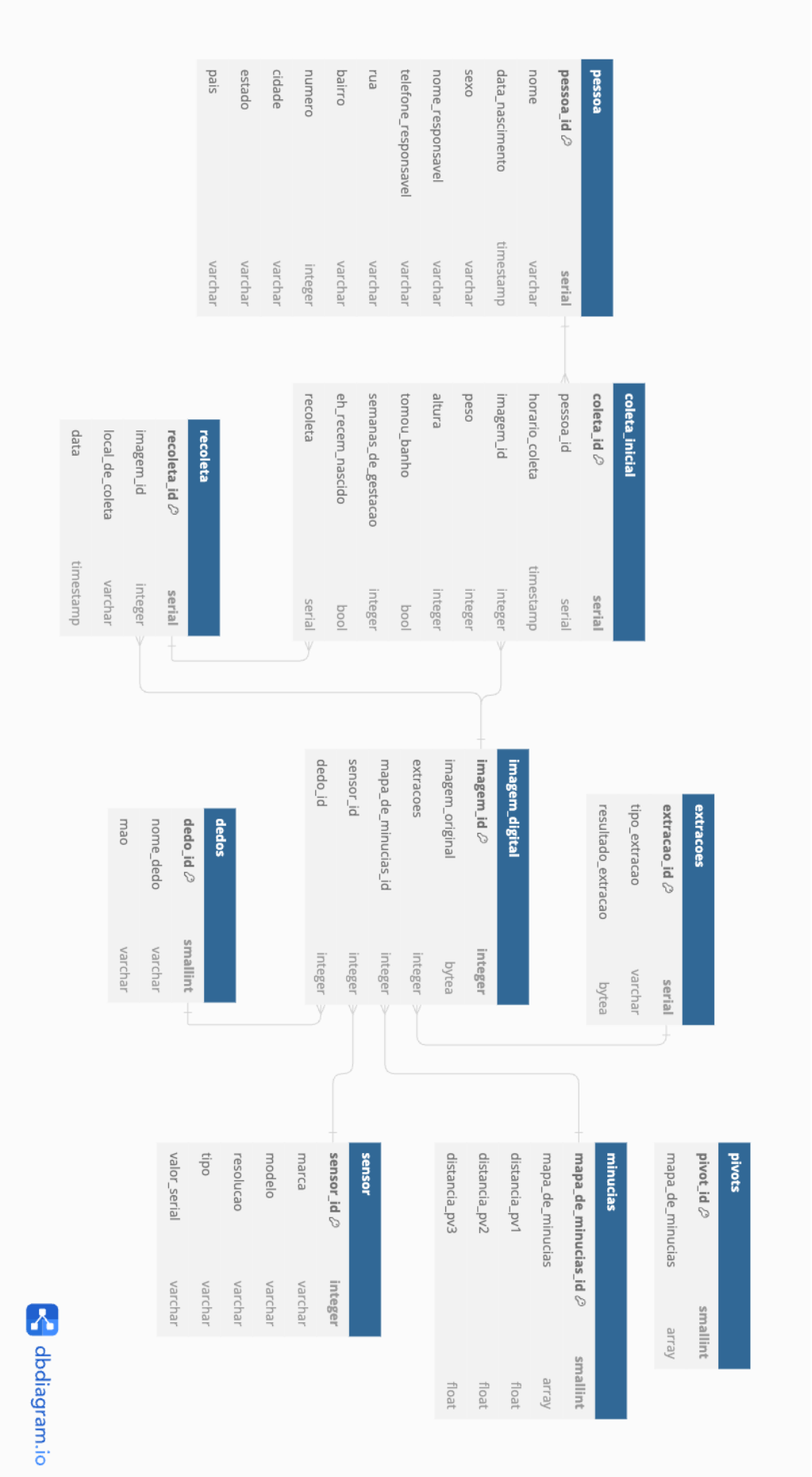
As tabelas "Minúcias" e "Pivots", como citado anteriormente, são as utilizadas pela técnica OMNI. Na tabela "Minúcias" são armazenados o mapa de minúcias da digital em questão e a distância desse mapa de minúcias para os focos 1, 2 e 3 da base de dados. É com esta tabela que a OMNI realiza a poda da base de dados, que foi exemplificada na seção 4.6. A tabela "Pivots" armazena os mapas de minúcias dos três focos comentados na seção 4.5.

5.2 Criação do banco de dados no Postgre

Com o modelo conceitual do banco já desenvolvido, foi criado o banco no PostgreSQL. No Apêndice A, está apresentado o código SQL utilizado para criar o banco de dados, seguindo o modelo estabelecido e suas respectivas relações. A seguir será apresentada uma comparação de exemplificação entre a base de dados original e a utilidade desta implementação.

Na base de dados original, os arquivos estão armazenados em pastas, sem qualquer identificação ou nomenclatura apropriada. Esse tipo de armazenamento dificulta buscas específicas. Por exemplo, caso um pesquisador esteja desenvolvendo um método de super-resolução

Figura 7 – Modelo conceitual do banco de dados



Fonte: Autoria própria (2024).

e precise de todas as imagens com resolução de 500 PPI capturadas no último ano, seria muito difícil encontrá-las nesse tipo de armazenamento.

Com a criação do banco e a inserção de dados de teste, pode ser feita uma consulta simples que pode ser utilizada para facilitar a busca de tipos específicos de digitais. Um exemplo de consulta está no apêndice B, e o resultado dessa consulta de teste está na figura 8.

Figura 8 – Resultado da consulta de exemplo

	pessoa_id integer	nome character varying (50)	horario_coleta timestamp without time zone	eh_recem_nascido boolean
1	1	Joao	2024-06-06 16:53:45.527108	false
2	2	Maria	2024-06-06 16:53:57.688229	false

Fonte: Autoria própria (2024).

Essa consulta foi realizada para digitais de teste no banco e serve apenas como exemplo. Com o desenvolvimento dos futuros projetos do grupo de pesquisa e a inserção de todas as digitais no banco de dados, será possível adaptar consultas para tipos diferentes de pesquisa, facilitando assim o desenvolvimento futuro.

5.3 Extração das características e população do banco de dados

Nesta etapa foi feito o tratamento dos dados e a extração das características necessárias para popular o banco de dados. Primeiramente foi feita a segmentação de todas as imagens da base de dados, e após isso foi extraído o mapa de minúcias, utilizando a biblioteca do grupo de pesquisa, todo esse processo foi feito de forma externa ao banco de dados. Como o resultado da extração é um arquivo com extensão xyt para cada digital da base de dados, foi necessário formatar e unificar todos os arquivos em um só, para facilitar a inserção no banco de dados. Com o arquivo formatado, foi utilizado o comando COPY do Postgre, comando este que copia todas as linhas de um arquivo e as insere como linhas de uma determinada tabela. Para as informações da pessoa foram escolhidos atributos clássicos, como nome, data de nascimento sexo, nome do responsável, endereço e telefone. Como na base de dados fornecida não se tem essas informações (se tem apenas a imagem da coleta), para os testes da OMNI apenas foram extraídos os mapas de minúcias. Para as futuras digitais que serão inseridas no banco com o avanço dos projetos do grupo de pesquisa, todas as informações necessárias já estão sendo coletadas para que a população do banco siga o modelo proposto. Os mapas de minúcias foram inseridos no formato de matriz, para facilitar os cálculos de distância da técnica e também por já estarem no formato esperado pelo *matching*.

5.4 Validação do MCC como medida de distância

Para demonstrar que o MCC respeita as 4 propriedades, foram selecionados 3 mapas de minúcias de digitais escolhidos aleatoriamente no banco de dados apenas para ilustrar os cálculos. Cada elemento dos mapas possui 4 dados: as coordenadas X e Y da minúcia na imagem, um valor booleano que indica se a minúcia é uma bifurcação ou não, e finalmente a angulação em radianos.

Para exemplificar, considerando a primeira minúcia do primeiro mapa, temos que a coordenada X é 52, a coordenada Y é 189, esta minúcia é uma bifurcação, e sua angulação em radianos é aproximadamente -1.518. Este é o formato em que todos os mapas estão armazenados no banco de dados.

Os mapas de minúcias selecionados para os testes são apresentados a seguir:

Mapa de minúcias da digital 1:

```
1  A = [(52, 189, True, -1.5182132651839548),
2             (102, 323, True, 3.141592653589793),
3             (30, 237, True, -2.0106389096106327),
4             (68, 279, True, 3.0890095919788516),
5             (211, 279, True, 2.539305307454829),
6             (161, 146, True, -0.9685089806599324),
7             (217, 62, True, -0.7853981633974483),
8             (186, 111, True, -0.8884797719201485),
9             (159, 141, True, 2.356194490192345),
10            (185, 249, True, 2.129395642138459)]
```

Mapa de minúcias da digital 2:

```
1  B = [(62, 296, True, -2.984990776607778),
2             (13, 236, True, 1.1999050379822342),
3             (234, 143, True, 2.129395642138459),
4             (138, 201, False, 2.386952733571283),
5             (47, 193, True, -1.6756732655251305),
6             (40, 143, True, 1.039072259536091),
7             (25, 39, True, -2.459276098715045),
8             (101, 176, False, 0.7146490776484106),
9             (176, 263, False, 2.335587608863185),
```

```
10              (43, 280, False, 0.6776739908510965)]
```

Mapa de minúcias da digital 3:

```
1  C = [(241, 161, True, -0.8567056281827387),
2               (202, 189, True, -1.0121970114513341),
3               (50, 186, True, 1.2490457723982544),
4               (167, 141, True, -0.9685089806599324),
5               (236, 148, True, -0.7483780475235183),
6               (135, 304, True, 2.701750070774057),
7               (158, 219, True, -1.039072259536091),
8               (95, 146, True, -2.173083672929861),
9               (190, 126, True, 2.2848870254070546),
10              (225, 271, True, 2.4668517113662407)]
```

A propriedade da Não-negatividade já é satisfeita pelo *matching*, visto que o resultado da distância sempre será um número entre 0 e 1, ou seja, maior ou igual a 0.

A identidade do indiscernível pede que a distância de um elemento para ele mesmo seja 0, o que no cálculo original do *matching* não é verdade, visto que o resultado de total similaridade entre dois elementos é 1. Para resolver isso foi feito um complemento de 1 no resultado da distância:

$$\text{Distância} = 1 - \text{Distância original do MCC} \tag{16}$$

Dessa maneira temos:

$$\text{Se: Distância original do MCC}(A, A) = 1.0 \tag{17}$$

$$\text{Então: } d(A, A) = 1.0 - 1.0 = 0.0 \tag{18}$$

em que d(A,A) representa a nova forma de calcular a distância aplicando o complemento de 1.

Assim a correspondência total entre dois elementos agora é dada pelo número 0, e a dissimilaridade total é dada pelo número 1, tornando assim válida a segunda propriedade.

Para a terceira propriedade foram realizadas medidas da distância do mapa de minúcias A para o mapa de minúcias B, e em seguida de B para A:

$$d(A, B) = 0.39 \tag{19}$$

$$d(B, A) = 0.39 \tag{20}$$

Como o *matching* das digitais não é interferido pela ordem dos elementos, o resultado foi o mesmo para os dois casos. Esse mesmo procedimento foi validado mais vezes no momento do cálculo das distâncias para os focos da OMNI, onde o *matching* respeitou a propriedade em todos os cenários testados.

Para a quarta propriedade foi feito um teste semelhante ao da terceira propriedade, onde foram medidas as distâncias entre os 3 mapas de minúcias citados:

$$d(A, B) = 0.39 \tag{21}$$

$$d(B, C) = 0.52 \tag{22}$$

$$d(A, C) = d(C, A) = 0.40 \tag{23}$$

Sendo assim:

$$d(A, C) \leq d(A, B) + d(B, C) \tag{24}$$

$$0.40 \leq 0.39 + 0.52 \tag{25}$$

E em todos os cenários testados o resultado das distâncias respeitou a quarta propriedade, fazendo assim com que o *matching*, em especial o MCC, possa ser utilizado como medida de distância métrica válida no contexto de digitais infantis.

5.5 Performance da técnica OMNI

Para a realização da análise sobre a performance da técnica OMNI foram realizadas quatro seções de testes de busca, onde uma digital aleatória era selecionada e então era feita a busca das digitais mais semelhantes com está de entrada. As quatro seções de teste foram as seguintes:

- 10 digitais mais semelhantes
- 5 digitais mais semelhantes
- 3 digitais mais semelhantes
- A digital mais semelhante

Ou seja, dada uma determinada digital de entrada, a ideia dos testes era buscar as mais semelhantes, que no caso seriam as possíveis candidatas para a correspondência. Um exemplo de aplicação nesse caso seria um hospital fazendo a conferência da digital de um recém-nascido no momento em que os pais estão levando a criança para casa. Seria coletada a digital do recém-nascido e então seria feita uma busca na base de dados para conferir se a

digital existe no banco de dados. Ou seja, é necessário buscar as digitais mais semelhantes à digital dessa criança para que possa ser liberada a saída do hospital. Os cenários de 10, 5, 3, e uma digital foram selecionados para simular variações da utilização da técnica e os resultados obtidos em cada uma serão apresentados a seguir. Como já dito anteriormente, a forma clássica de se buscar uma digital no banco é por meio de uma busca exaustiva, onde o *matching* é feito entre a digital de entrada e todas as digitais do banco, buscando assim a mais semelhante. Dessa forma, para que possa ser analisado o desempenho da OMNI, primeiro deve ser feita a busca sequencial na base de dados. A busca sequencial foi feita utilizando um comando *select* simples junto do MCC.

A busca sequencial manteve um desempenho semelhante buscando as 10, 5, 3 ou uma digital mais parecida visto que o processo é sempre o mesmo: percorrer toda a base de dados e ordenar por semelhança. O tempo médio da busca sequencial foi de 6 minutos e 20 segundos, com desvio padrão de 18,4 segundos.

A seguir serão apresentados os resultados da OMNI.

5.5.1 Redução do espaço de busca

Todos os testes foram feitos entre raios de busca de 0.020 e 0.060 com variação de 0.005 entre cada raio. O raio é importante pois é ele quem define a redução do espaço de busca. Esses valores foram escolhidos pois representam um arredondamento da maior e da menor distância da base utilizada. A tabela 3 apresenta essa redução de busca para cada um dos raios utilizados nos testes.

Tabela 3 – Redução do espaço de busca para cada raio

Raio de busca	Núm. elementos comparados	Redução do espaço de busca (%)
0.020	127	94,40
0.025	212	92,31
0.030	365	86,77
0.035	536	80,57
0.040	744	73,02
0.045	984	64,32
0.050	1214	55,98
0.055	1427	48,26
0.060	1636	40,68

Fonte: Autoria própria (2024).

A tabela 4 apresenta o tempo médio e o desvio padrão para cada um dos raios utilizados nos testes. O maior desvio padrão foi de 7,1 segundos e o menor foi de 0,7 segundos.

Embora raios menores representem uma maior poda do espaço de busca, vale destacar que isso diretamente restringe as possibilidades de acerto da digital que está sendo buscada, visto que como são poucos elementos, pequenas variações do resultado do MCC podem significar que a digital procurada acabe ficando de fora do espaço de busca. Desta forma, foram

Tabela 4 – Tempo médio e desvio padrão para cada raio

Raio de busca	Tempo médio (m:s)	Desvio padrão (s)
0.020	00:19	0,7
0.025	00:33	1,9
0.030	00:59	1,6
0.035	01:30	7,1
0.040	02:03	2,6
0.045	02:43	5,7
0.050	03:16	5,3
0.055	03:41	5,0
0.060	04:06	5,0

Fonte: Autoria própria (2024).

conduzidos testes onde seriam buscadas as N digitais mais semelhantes a uma determinada digital, e os resultados destes testes são apresentados a seguir.

5.5.2 Buscando os 10 elementos mais semelhantes com uma determinada digital

A tabela 5 apresenta os resultados da técnica OMNI para os 10 elementos mais semelhantes.

Tabela 5 – Resultados da busca para as 10 digitais mais semelhantes

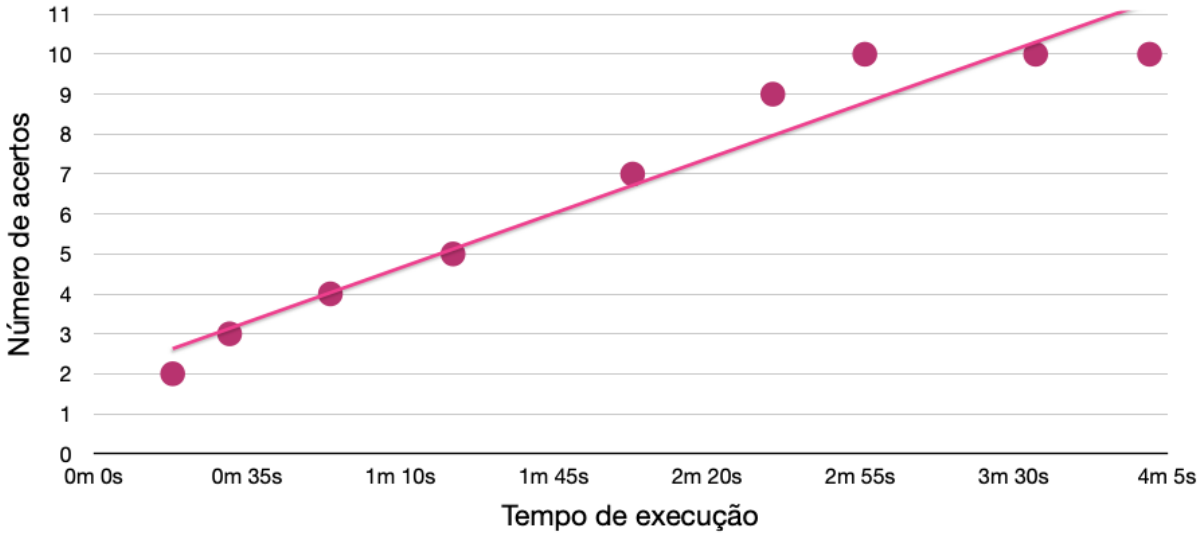
Raio de busca	Núm. acertos	Tempo (m:s)
0.020	2	0:18
0.025	3	0:31
0.030	4	0:54
0.035	5	1:22
0.040	7	2:03
0.045	9	2:35
0.050	10	2:56
0.055	10	3:35
0.060	10	4:01

Fonte: Autoria própria (2024).

Nota-se que a partir de um raio de busca de 0.050 a OMNI foi capaz de encontrar os 10 elementos mais semelhantes com uma taxa de acerto de 100%. Com esse raio de busca o número de elementos comparados representa uma redução de aproximadamente 56% do espaço de busca. Ao aplicar a busca sequencial nesse novo espaço de busca, é obtido o tempo de 2 minutos e 56 segundos, ou seja, cortando praticamente pela metade o tempo de busca necessário para encontrar os 10 elementos mais semelhantes. Vale destacar que nesse tempo de execução apresentado na tabela estão inclusos o tempo necessário para a OMNI fazer a poda e também a busca sequencial.

O gráfico 9 apresenta o número de acertos por tempo de busca, e também apresenta a linha de tendência para a busca das 10 digitais mais semelhantes.

Figura 9 – Gráfico de número de acertos por tempo de busca para 10 digitais



Fonte: Autoria própria (2024).

5.5.3 Buscando os 5 elementos mais semelhantes com uma determinada digital

A tabela 6 apresenta os resultados da técnica OMNI para os 5 elementos mais semelhantes.

Tabela 6 – Resultados da busca para as 5 digitais mais semelhantes

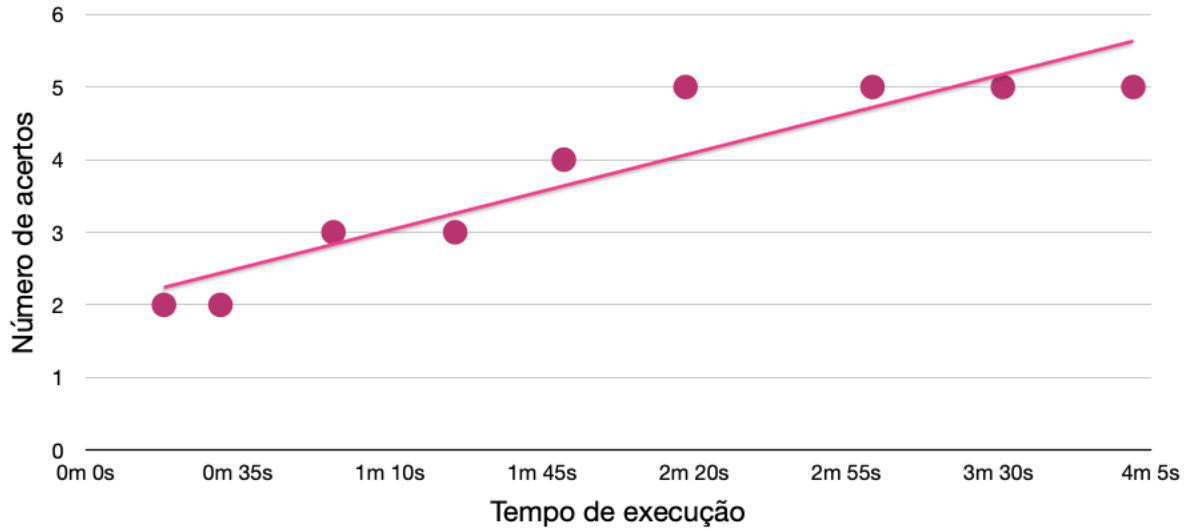
Raio de busca	Núm. acertos	Tempo (m:s)
0.020	2	0:18
0.025	2	0:31
0.030	3	0:57
0.035	3	1:25
0.040	4	1:50
0.045	5	2:18
0.050	5	3:01
0.055	5	3:31
0.060	5	4:01

Fonte: Autoria própria (2024).

Para encontrar as 5 digitais mais semelhantes foi necessário um raio de 0.045, representando uma redução do espaço de busca de aproximadamente 64% e também um tempo de busca de 2 minutos e 18 segundos (também incluso o tempo de poda).

O gráfico 10 apresenta o número de acertos por tempo de busca, e também apresenta a linha de tendência para a busca das 5 digitais mais semelhantes.

Figura 10 – Gráfico de número de acertos por tempo de busca para 5 digitais



Fonte: Autoria própria (2024).

5.5.4 Buscando os 3 elementos mais semelhantes com uma determinada digital

A tabela 7 apresenta os resultados da técnica OMNI para os 3 elementos mais semelhantes.

Tabela 7 – Resultados da busca para as 3 digitais mais semelhantes

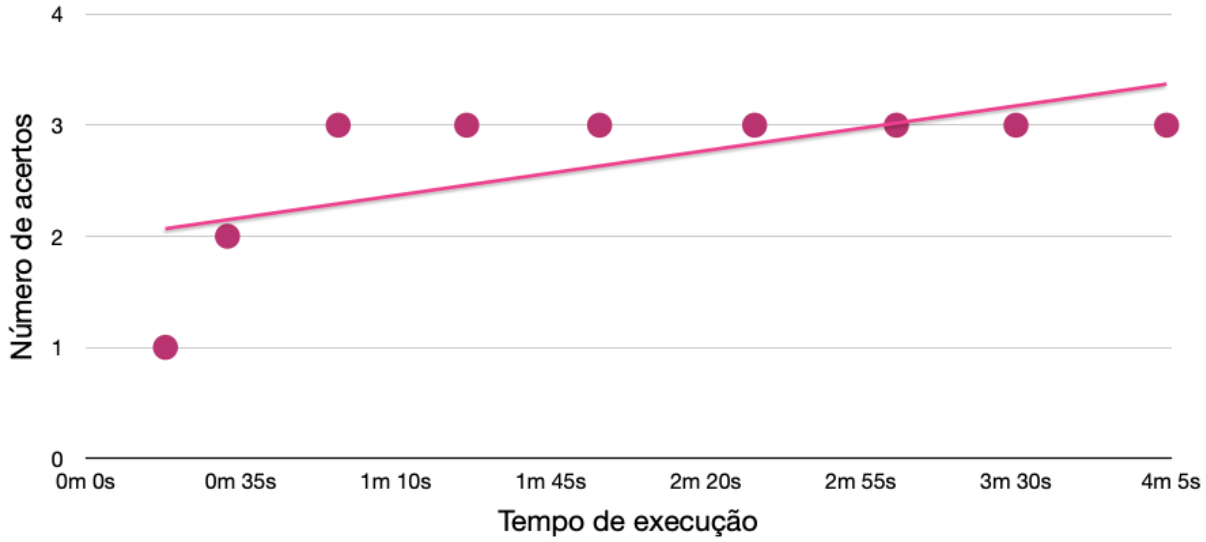
Raio de busca	Núm. acertos	Tempo (m:s)
0.020	1	0:18
0.025	2	0:32
0.030	3	0:57
0.035	3	1:26
0.040	3	1:56
0.045	3	2:31
0.050	3	3:03
0.055	3	3:30
0.060	3	4:04

Fonte: Autoria própria (2024).

Para encontrar as 3 digitais mais semelhantes foi necessário um raio de 0.030. Este raio apresenta uma redução do espaço de busca de aproximadamente 87% e um tempo de busca médio de 57 segundos (também incluso o tempo da poda). Nesse caso já temos um cenário que pode ser utilizado como margem de erro, ou seja, um cenário onde a poda é significativa, o tempo de busca é consideravelmente reduzido e ainda assim foram encontradas três digitais candidatas a serem correspondentes ao indivíduo que está sendo buscado.

O gráfico 11 apresenta o número de acertos por tempo de busca, e também apresenta a linha de tendência para a busca das 3 digitais mais semelhantes.

Figura 11 – Gráfico de número de acertos por tempo de busca para 3 digitais



Fonte: Autoria própria (2024).

5.5.5 Buscando a digital mais semelhante com a digital de entrada

A tabela 8 apresenta os resultados da técnica OMNI para buscar o elemento mais semelhante à digital de entrada. Este cenário apresenta um resultado mais real sobre a técnica, e também o resultado mais importante, visto que em uma busca em larga escala normalmente será buscado o elemento mais semelhante de todos.

Tabela 8 – Resultados da busca da digital mais semelhante

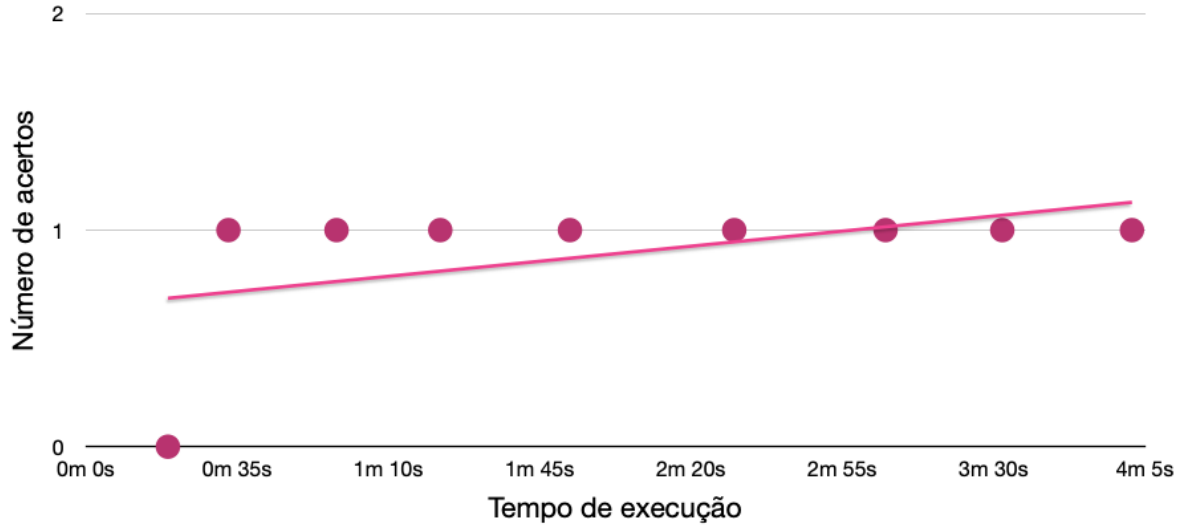
Raio de busca	Núm. acertos	Tempo (m:s)
0.020	0	0:19
0.025	1	0:33
0.030	1	0:58
0.035	1	1:22
0.040	1	1:52
0.045	1	2:30
0.050	1	3:05
0.055	1	3:32
0.060	1	4:02

Fonte: Autoria própria (2024).

Nota-se que a partir de um raio de 0.025 já foi possível encontrar a digital mais semelhante à digital de entrada. Esse resultado apresenta um espaço de busca de apenas 212 elementos, com uma redução de aproximadamente 92% do tamanho original da base de dados. Com esse raio, foi possível realizar a busca em apenas 33 segundos, uma melhoria significativa em comparação aos 6 minutos da busca sequencial clássica.

O gráfico 12 apresenta o número de acertos por tempo de busca nesse cenário.

Figura 12 – Gráfico de número de acertos por tempo de busca para a digital mais semelhante



Fonte: Autoria própria (2024).

Devido ao tamanho pequeno da base de dados em questão, a utilização de índices na consulta não representou muita melhoria e foi desconsiderada. Todos os resultados apresentados aqui foram testados na mesma máquina sob as mesmas condições.

6 CONCLUSÃO

O presente trabalho teve como objetivo geral a validação da eficácia da técnica OMNI em um banco de dados dedicado à biometria digital infantil, o que foi comprovado pelos resultados apresentados. A técnica demonstrou ser eficiente na redução do espaço de busca e, principalmente, manteve o desempenho esperado, mesmo em um banco de dados composto apenas por digitais infantis. A redução de 92% do espaço de busca ao buscar a digital mais semelhante com a entrada é um resultado muito significativo, principalmente se colocado em escala com o tamanho que bancos de dados desse tipo podem atingir. Além disso, foi proposto verificar a utilização do *matching* como medida de distância em um espaço métrico, que, por sua vez, apresentou resultados promissores, especialmente pelo fato de que manteve a técnica OMNI em funcionamento e consolidou os cálculos de distância necessários.

Vale destacar que, além de um resultado determinístico, dependente apenas do tamanho da base de dados e do número de minúcias comparadas, o principal resultado é a poda em si, ou seja, a porcentagem de redução do espaço de busca. Com um espaço de busca reduzido, é possível implementar outras técnicas de busca, como, por exemplo, um algoritmo de KNN ou busca utilizando GPUs, resultando em tempos possivelmente melhores.

Outra observação importante é que os resultados da OMNI são dependentes da função de distância, nesse caso, o *matching*. Como o *matching* de digitais infantis ainda apresenta dificuldades na literatura atual, e considerando que todos os cálculos deste projeto foram realizados em digitais cuja qualidade não era garantida, esses resultados podem ser considerados como obtidos em um dos piores cenários possíveis. Portanto, à medida que melhorias forem implementadas no *matching* e incorporadas ao banco de dados, os resultados deste trabalho tendem a se concretizar e melhorar ainda mais.

A principal vantagem do método proposto, além da redução do espaço de busca, é o potencial para futuras melhorias e experimentos. Este trabalho pode inspirar muitos outros, como, por exemplo, a implementação de outros métodos de busca na base reduzida criada pela OMNI, conforme citado anteriormente, melhorias no *matching* conforme novos estudos surgirem. Também pode ser interessante uma análise sobre os parâmetros escolhidos para a OMNI e para o MCC, e como essas mudanças de parâmetros afetam o resultado final da técnica, além disso, poderia ser feito um estudo com um aumento na base de dados para ver o quanto isso afeta o desempenho.

No artigo original da OMNI, os resultados foram de uma redução do tempo de busca de até 10 vezes, o que se manteve nos experimentos realizados neste trabalho. Também é importante ressaltar que uma das principais qualidades da OMNI é a sua escalabilidade, sendo assim a OMNI tende a manter uma poda linear com o aumento da base de dados, então é esperado que com o aumento da base de dados do grupo de pesquisa, os resultados de redução do espaço de busca aqui apresentados se mantenham. Sendo assim, pode-se afirmar que a

técnica OMNI é uma boa escolha para a redução do espaço de busca para digitais infantis e que o *matching* é uma medida válida de distância métrica para este tipo de dados.

REFERÊNCIAS

ALBANEZ MARCOS AURéLIO BATISTA, S. F. d. S. Daniela de O. O que é segmentação de imagens? 2016.

BASAK, P. *et al.* Multimodal biometric recognition for toddlers and pre-school children. *In*: IEEE. **2017 IEEE International Joint Conference on Biometrics (IJCB)**. [*S.l.*], 2017. p. 627–633.

BAZEN, A.; GEREZ, S. Systematic methods for the computation of the directional fields and singular points of fingerprints. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 24, n. 7, p. 905–919, 2002.

BOER, J. D.; BAZEN, A. M.; GEREZ, S. H. Indexing fingerprint databases based on multiple features. *In*: **ProRISC the 12th Annual Workshop on Circuits, Systems and Signal Processing**. [*S.l.*: *s.n.*], 2001. p. 300–306.

CAPPELLI, R.; FERRARA, M.; MALTONI, D. Minutia cylinder-code: A new representation and matching technique for fingerprint recognition. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 32, n. 12, p. 2128–2141, 2010.

CAPPELLI, R.; FERRARA, M.; MALTONI, D. Large-scale fingerprint identification on gpu. **Information Sciences**, v. 306, p. 1–20, 2015. ISSN 0020-0255. Disponível em: https://www.sciencedirect.com/science/article/pii/S0020025515001097.

CHEN, J.-Y.; BOUMAN, C.; ALLEBACH, J. Fast image database search using tree-structured vq. *In*: **Proceedings of International Conference on Image Processing**. [*S.l.*: *s.n.*], 1997. v. 2, p. 827–830 vol.2.

ENGELSMA, J. J. *et al.* Infant-id: Fingerprints for global good. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, IEEE, v. 44, n. 7, p. 3543–3559, 2021.

ENGELSMA, J. J. *et al.* Infant-prints: Fingerprints for reducing infant mortality. *In*: **Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops**. [*S.l.*: *s.n.*], 2019. p. 67–74.

ESPINOSA-DURO, V. Minutiae detection algorithm for fingerprint recognition. **IEEE Aerospace and Electronic Systems Magazine**, v. 17, n. 3, p. 7–10, 2002.

GHOSH, D.; CABRERA, J. Enriched random forest for high dimensional genomic data. **IEEE/ACM Transactions on Computational Biology and Bioinformatics**, v. 19, n. 5, p. 2817–2828, 2022.

ILOANUSI, O. N.; OSUAGWU, C. C. Clustering: Applied to data structuring and retrieval. **International Journal of Advanced Computer Science and Applications**, Citeseer, v. 2, n. 11, 2011.

INDRAWAN, G. I. G.; SITOHANG, B. S. B.; AKBAR, S. A. S. Review of sequential access method for fingerprint identification. **TELKOMNIKA (Telecommunication Computing Electronics and Control)**, v. 10, n. 2, p. 335–342, 2012.

JAIN, A. K.; ROSS, A. A.; NANDAKUMAR, K. **Introduction to Biometrics**. Springer US, 2011. Disponível em: https://doi.org/10.1007/978-0-387-77326-1.

LIMA, E. L. **Espaços métricos**. [*S.l.*]: IMPA, 2020.

LUMINI, A.; MAIO, D.; MALTONI, D. Continuous versus exclusive classification for fingerprint retrieval. **Pattern Recognition Letters**, v. 18, n. 10, p. 1027–1034, 1997. ISSN 0167-8655. Disponível em: https://www.sciencedirect.com/science/article/pii/S016786559700127X.

MAGALHãES, H. D. S. P. S. **Biometria e autenticação**. [*S.l.*]: Associação Portuguesa de Sistemas de Informação (APSI), 2003.

MATSUI, C. J. M. **Consultas por similaridade em bases de dados complexos utilizando técnica OMNI em SGBDR**. 2018. Dissertação (B.S. thesis) — Universidade Tecnológica Federal do Paraná, 2018.

MEHROTRA, H. *et al.* Feature level clustering of large biometric database. **arXiv preprint arXiv:1002.0383**, 2010.

MHATRE, A. J. *et al.* Efficient search and retrieval in biometric databases. *In*: SPIE. **Biometric technology for human identification II**. [*S.l.*], 2005. v. 5779, p. 265–273.

MOHN, E. **SQL (Structured Query Language)**. [*S.l.*]: Salem Press, 2023. Encyclopedia of Science, Research Starters.

NATOSAFE. **Quais são os tipos de impressões digitais?** 2022. Disponível em: https://natosafe.com.br/tipos-de-impressao-digital/. Acesso em: 30 out. 2023.

Organização das Nações Unidas. **Global Report on Trafficking in Persons 2022**. 2022. Disponível em: https://www.unodc.org/documents/data-and-analysis/glotip/2022/GLOTiP_2022_web.pdf. Acesso em: 22 ago. 2023.

POLA, I. R. V. **Explorando conceitos da teoria de espaços métricos em consultas por similaridade sobre dados complexos**. 2010. Tese (Tese de Doutorado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2010. Disponível em: Acessoem:2023-10-31.

POSTGRESQL. **PostgreSQL Documentation**. 2023. Acesso em: 13 de Novembro de 2023. Disponível em: https://www.postgresql.org/docs/current/tutorial-concepts.html.

POSTGRESQL. **PostgreSQL: The world's most advanced open source database**. 2023. Disponível em: https://www.postgresql.org/.

PRECIOZZI, J. *et al.* Fingerprint biometrics from newborn to adult: A study from a national identity database system. **IEEE Transactions on Biometrics, Behavior, and Identity Science**, v. 2, n. 1, p. 68–79, 2020.

PRIMO, J. J. B. *et al.* Métodos para extração de atributos em imagens de impressão digital. Universidade Federal de Campina Grande, 2019.

RATHA, N. K.; CHEN, S.; JAIN, A. K. Adaptive flow orientation-based feature extraction in fingerprint images. **Pattern Recognition**, v. 28, n. 11, p. 1657–1672, 1995. ISSN 0031-3203. Disponível em: https://www.sciencedirect.com/science/article/pii/0031320395000393.

SOUTHIER, L. F. P. *et al.* **Systematic Literature Review on Neonatal Fingerprint Recognition**. 2023. Preprint.

TRAINA, A. *et al.* Similarity search without tears: the omni-family of all-purpose access methods. *In*: IEEE. **Proceedings 17th International Conference on Data Engineering**. [*S.l.*], 2001. p. 623–630.

TRAINA, C. *et al.* The omni-family of all-purpose access methods: a simple and effective way to make similarity search more efficient. **VLDB**, 2006.

WANG, Z.; CHEN, J.; HOI, S. C. H. Deep learning for image super-resolution: A survey. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 43, n. 10, p. 3365–3387, 2021.

WU, H. *et al.* Face recognition based on convolution siamese networks. *In*: **2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)**. [*S.l.*: *s.n.*], 2017. p. 1–5.

APÊNDICES

APÊNDICE A – Código de Criação do Banco de Dados

```
1  -- Table: Pivos
2  CREATE TABLE pivots (
3      id SERIAL PRIMARY KEY,
4      vetor_minucias float[][]
5  );
6
7  --Table: Minucias
8  CREATE TABLE minucias (
9      id SERIAL PRIMARY KEY,
10     vetor_minucias float[][],
11    distancepivot1 float,
12    distancepivot2 float,
13    distancepivot3 float
14 )
15
16 -- Table: pessoa
17 CREATE TABLE pessoa (
18     pessoa_id SERIAL PRIMARY KEY,
19     nome VARCHAR(50),
20     data_nascimento TIMESTAMP,
21     sexo VARCHAR(15),
22     nome_responsavel VARCHAR(50),
23     telefone_responsavel VARCHAR(20),
24     rua VARCHAR(50),
25     bairro VARCHAR(50),
26     numero INTEGER,
27     cidade VARCHAR(50),
28     estado VARCHAR(25),
29     pais VARCHAR(30)
30 );
31
32 -- Table: dedos
33 CREATE TABLE dedos (
34     dedo_id SMALLINT PRIMARY KEY,
```

```
35      nome_dedo VARCHAR(15),
36      mao VARCHAR(10)
37  );
38
39  -- Table: extracoes
40  CREATE TABLE extracoes (
41      extracao_id SERIAL PRIMARY KEY,
42      tipo_extracao VARCHAR(25),
43      resultado_extracao BYTEA
44  );
45
46  -- Table: sensor
47  CREATE TABLE sensor (
48      sensor_id INTEGER PRIMARY KEY,
49      marca VARCHAR(30),
50      modelo VARCHAR(30),
51      resolucao VARCHAR(30),
52      tipo VARCHAR(50),
53      valor_serial VARCHAR(50)
54  );
55
56  -- Table: imagem_digital
57  CREATE TABLE imagem_digital (
58      imagem_id INTEGER PRIMARY KEY,
59      imagem_original BYTEA,
60      extracoes INTEGER REFERENCES extracoes(extracao_id),
61      mapa_de_minucias_id INTEGER REFERENCES minucias(id),
62      sensor_id INTEGER REFERENCES sensor(sensor_id),
63      dedo_id INTEGER REFERENCES dedos(dedo_id)
64  );
65
66  -- Table: recoleta
67  CREATE TABLE recoleta (
68      recoleta_id SERIAL PRIMARY KEY,
```

```
69      imagem_id INTEGER REFERENCES imagem_digital(imagem_id),
70      local_de_coleta VARCHAR(50),
71      data TIMESTAMP
72  );
73
74  -- Table: coleta_inicial
75  CREATE TABLE coleta_inicial (
76      coleta_id SERIAL PRIMARY KEY,
77      pessoa_id INTEGER REFERENCES pessoa(pessoa_id),
78      horario_coleta TIMESTAMP,
79      imagem_id INTEGER REFERENCES imagem_digital(imagem_id),
80      peso INTEGER,
81      altura INTEGER,
82      tomou_banho BOOLEAN,
83      semanas_de_gestacao INTEGER,
84      eh_recem_nascido BOOLEAN,
85      recoleta INTEGER REFERENCES recoleta(recoleta_id)
86  );
```

APÊNDICE B – Código de exemplo de consulta

```
1  SELECT
2      p.pessoa_id,
3      p.nome,
4    ci.horario_coleta,
5    ci.eh_recem_nascido
6  FROM
7      pessoa p
8  JOIN
9      coleta_inicial ci ON p.pessoa_id = ci.pessoa_id
10 JOIN
11     imagem_digital id ON ci.imagem_id = id.imagem_id
12 JOIN
13     sensor s ON id.sensor_id = s.sensor_id
14 WHERE
15     s.resolucao = '500';
```

APÊNDICE C – Código do MCC no banco de dados

```
1  CREATE OR REPLACE FUNCTION calc_dist_between_two_maps(
2  vetor_minucias_1 float[][], vetor_minucias_2 float[][])
3  RETURNS float
4  LANGUAGE plpython3u
5  AS $$
6      import numpy as np
7      import math
8
9      mcc_sigma_s = 28.0/3.0
10     mcc_tau_psi = 400.0
11     mcc_mu_psi = 0.01
12
13     def sigmoid(v,u,t):
14         expo = math.exp(-t * (v-u))
15         return 1 / (1+expo)
16
17     def Gs(t_sqr):
18         return np.exp(-0.5 * t_sqr / (mcc_sigma_s**2)) /
19         (math.tau**0.5 * mcc_sigma_s)
20
21     def Psi(v):
22         return 1.0/(1.0 + np.exp(-mcc_tau_psi * (v - mcc_mu_psi)))
23
24     mcc_radius = 70
25     mcc_size = 16
26
27     g = 2 * mcc_radius / mcc_size
28     x = np.arange(mcc_size)*g - (mcc_size/2)*g + g/2
29     y = x[..., np.newaxis]
30     iy, ix = np.nonzero(x**2 + y**2 <= mcc_radius**2)
31     ref_cell_coords = np.column_stack((x[ix], x[iy]))
32
33     xyd = np.array([(x,y,d) for x,y,_,d in vetor_minucias_1])
34     d_cos = np.cos(xyd[:,2]).reshape((-1,1,1))
```

```
35      d_sin = np.sin(xyd[:,2]).reshape((-1,1,1))
36      rot = np.block([[d_cos, d_sin], [-d_sin, d_cos]])
37      xy = xyd[:,:2]
38      cell_coords = np.transpose(rot@ref_cell_coords.T +
39      xy[:,:,np.newaxis],[0,2,1])
40      dists = np.sum((cell_coords[:,:,np.newaxis,:] - xy)**2, -1)
41      cs = Gs(dists)
42      diag_indices = np.arange(cs.shape[0])
43      cs[diag_indices,:,diag_indices] = 0
44      local_structures = Psi(np.sum(cs, -1))
45
46      # Calculando as distancias
47      xyd2 = np.array([(x,y,d) for x,y,_,d in vetor_minucias_2])
48      d_cos2, d_sin2 = np.cos(xyd2[:,2]).reshape((-1,1,1)),
49      np.sin(xyd2[:,2]).reshape((-1,1,1))
50      rot2 = np.block([[d_cos2, d_sin2], [-d_sin2, d_cos2]])
51
52      xy2 = xyd2[:,:2]
53      cell_coords2 = np.transpose(rot2@ref_cell_coords.T +
54      xy2[:,:,np.newaxis],[0,2,1])
55      dists2 = np.sum((cell_coords2[:,:,np.newaxis,:] - xy2)**2, -1)
56
57      cs2 = Gs(dists2)
58      diag_indices2 = np.arange(cs2.shape[0])
59      cs2[diag_indices2,:,diag_indices2] = 0
60
61      local_structures2 = Psi(np.sum(cs2, -1))
62
63      # Calculando as distancias finais
64      dists = np.sqrt(np.sum((local_structures[:,np.newaxis,:] -
65      local_structures2)**2, -1))
66      dists /= (np.sqrt(np.sum(local_structures**2, 1))[:,np.newaxis] +
67      np.sqrt(np.sum(local_structures2**2, 1)))
68
```

```
69      minNp = 4
70      maxNp = 12
71      up = 20
72      tp = 2/5
73      Na = len(vetor_minucias_1)
74      Nb = len(vetor_minucias_2)
75      num_p = minNp + round(sigmoid(min(Na,Nb), up, tp) * (maxNp-minNp))
76      pairs = np.unravel_index(np.argpartition(dists, num_p, None)[:num_p],
77      dists.shape)
78      score =  np.mean(dists[pairs[0], pairs[1]])
79
80      return score
81  $$;
```

APÊNDICE D – Código da técnica OMNI no banco de dados

```
        ]
```

```
1  CREATE OR REPLACE FUNCTION OMNI1PvRangeQuery(qc float[][], radius float,
2  max_score float)
3  returns table (id integer, distance float) as $$
4  DECLARE
5      distance_id1 FLOAT;
6      distance_id2 FLOAT;
7      distance_id3 FLOAT;
8      calc_dist FLOAT;
9  BEGIN
10     -- Calcule as distancias dos pivos
11     SELECT
12         calc_dist_between_two_maps(
13             (qc),
14             (SELECT vetor_minucias FROM pivots WHERE pivots.id = 1)
15         ) INTO distance_id1;
16
17     SELECT
18         calc_dist_between_two_maps(
19             (qc),
20             (SELECT vetor_minucias FROM pivots WHERE pivots.id = 2)
21         ) INTO distance_id2;
22
23     SELECT
24         calc_dist_between_two_maps(
25             (qc),
26             (SELECT vetor_minucias FROM pivots WHERE pivots.id = 3)
27         ) INTO distance_id3;
28
29   RETURN QUERY
30   SELECT falsepositives.id, calc_dist_between_two_maps(
31       (qc),
32       (falsepositives.vetor_minucias)
33     )
```

```
34    FROM (
35      SELECT minucias.id, minucias.vetor_minucias
36      FROM minucias
37      WHERE distancepivot1 BETWEEN distance_id1 - radius AND distance_id1 + radius
38        AND distancepivot2 BETWEEN distance_id2 - radius AND distance_id2 + radius
39        AND distancepivot3 BETWEEN distance_id3 - radius AND distance_id3 + radius
40      GROUP BY minucias.id
41    ) AS falsepositives
42    WHERE calc_dist_between_two_maps(
43      (qc),
44      (falsepositives.vetor_minucias)
45    ) <= max_score
46    ORDER BY calc_dist_between_two_maps(
47      (qc),
48      (falsepositives.vetor_minucias)
49    );
50  END $$
```