MINISTÉRIO DA EDUCAÇÃO

SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA BAIANO - CAMPUS CATU

WILLIAM SCHRAMM BARROS

TRANSFORMAÇÃO DIGITAL NO CADASTRAMENTO ACADÊMICO: Uso de Processamento Digital de Imagens e OCR em documentos oficiais no IF Baiano – Campus Catu

CATU-BA

2026

WILLIAM SCHRAMM BARROS

TRANSFORMAÇÃO DIGITAL NO CADASTRAMENTO ACADÊMICO: Uso de Processamento Digital de Imagens e OCR em documentos oficiais no IF Baiano – Campus Catu

Trabalho de conclusão de curso apresentado ao curso de Análise e Desenvolvimento de Sistemas como requisito parcial para obtenção do grau de Tecnólogo em Análise e Desenvolvimento de Sistemas pelo Instituto Federal de Educação, Ciência e Tecnologia Baiano - Campus Catu.
Orientador: Prof. Dr. André Luiz Andrade Rezende
Coorientador: Prof. Me. Cayo Pabllo Santana de Jesus

CATU-BA

2026

Instituto Federal de Educação, Ciência e Tecnologia Baiano – Campus Catu Setor de Biblioteca

B277 Barros, William Schramm

Transformação digital no cadastramento acadêmico: uso de processamento digital e imagens e OCR em documentos oficiais no IF Baiano – Campus Catu / William Schramm. – 2026.

82 f.: il. color.

Orientador(a): Prof. Dr. André Luiz Andrade Rezende.
Co-orientador(a): Prof. Ms. Cayo Pabllo Santana de Jesus.

TCC (monografia), Instituto Federal de Educação, Ciência e Tecnologia Baiano, Tecnólogo em Análise e Desenvolvimento de Sistemas, Catu, 2026.

1. Processamento digital de imagens. 2. Reconhecimento óptico de caracteres. 3. documentos oficiais. I. Rezende, André Luiz Andrade. II. Jesus, Cayo Pabllo Santana de . III. Título.

CDU: 004.93

WILLIAM SCHRAMM BARROS

TRANSFORMAÇÃO DIGITAL NO CADASTRAMENTO ACADÊMICO: Uso de Processamento Digital de Imagens e OCR em documentos oficiais no IF Baiano – Campus Catu

Trabalho de conclusão de curso apresentado ao curso de Análise e Desenvolvimento de Sistemas como requisito parcial para obtenção do grau de Tecnólogo em Análise e Desenvolvimento de Sistemas pelo Instituto Federal de Educação, Ciência e Tecnologia Baiano - Campus Catu.
Orientador: Prof. Dr. André Luiz Andrade Rezende
Co-orientador: Prof. Me. Cayo Pabllo Santana de Jesus

Data de Aprovação: _____ /_____ /_____.

_____

Prof. Dr. André Luiz Andrade Rezende – Presidente

(Orientador – Instituto Federal Baiano – Campus Catu)

_____

Prof. Dr. Romero Mendes Freire de Moura Junior – Membro

(Banca examinadora – Instituto Federal Baiano – Campus Catu)

_____

Prof. Dr. Társio Ribeiro Cavalcante – Membro

(Banca examinadora – Instituto Federal Baiano – Campus Catu)

RESUMO

Este trabalho apresenta o desenvolvimento de um modelo computacional voltado a apoiar o processo de matrícula e cadastramento de estudantes na Secretaria de Registros Acadêmicos (SRA) do IF Baiano — Campus Catu, mediante a identificação, a extração e a organização automatizadas de informações presentes no Registro Geral (RG), na Carteira de Identidade Nacional (CIN) e na Carteira Nacional de Habilitação (CNH), com validação assistida pelo usuário. A pesquisa foi motivada pela necessidade de apoiar atividades administrativas que dependem da consulta, conferência e digitação manual de dados documentais no Sistema Unificado de Administração Pública (SUAP). Para isso, foi desenvolvido um protótipo utilizando técnicas de Processamento Digital de Imagens, Reconhecimento Óptico de Caracteres, expressões regulares, regras documentais, telas de validação assistida pelo usuário, exportação dos dados em arquivo de valores separados por vírgula (CSV) e registro de eventos em logs. A avaliação foi realizada em ambiente laboratorial controlado, sem integração ao SUAP e sem validação operacional com os servidores da SRA, utilizando 30 documentos distribuídos entre RG, CIN e CNH, incluindo amostras documentais obtidas por diferentes meios e documentos sintéticos. Os resultados foram classificados como completos, parciais ou falhos, permitindo calcular a Taxa de Acerto (TA) e a Taxa de Reconhecimento Aproveitável (TRA). A TA, que considerou somente os campos integralmente extraídos, foi de 38,6% para o RG, 67,5% para a CIN e 70,0% para a CNH. Como indicador complementar, a TRA, que reuniu resultados completos e parciais, foi de 75,7%, 78,8% e 80,0%, respectivamente. Considerando os três tipos documentais, a TA geral foi de 60,4% e a TRA geral foi de 78,4%. Os resultados parciais permaneceram dependentes de conferência e correção manual pelo usuário. Conclui-se que a solução proposta apresentou potencial de apoio à extração inicial e à organização de informações documentais em ambiente laboratorial, mas ainda necessita de validação em ambiente real da SRA, testes com usuários, ampliação da base experimental e avaliação de eventual integração segura com o SUAP.

Palavras-chave: processamento digital de imagens; reconhecimento óptico de caracteres; documentos oficiais; cadastramento acadêmico; transformação digital.

ABSTRACT

This work presents the development of a computational model designed to support student enrollment and academic registration activities at the Secretaria de Registros Acadêmicos (SRA), the Academic Records Office of the Instituto Federal de Educação, Ciência e Tecnologia Baiano (IF Baiano) — Campus Catu, through the automated identification, extraction, and organization of information contained in the Brazilian identity documents Registro Geral (RG), Carteira de Identidade Nacional (CIN), and Carteira Nacional de Habilitação (CNH), with user-assisted validation. The research was motivated by the need to support administrative activities that depend on the consultation, verification, and manual entry of documentary data into the Sistema Unificado de Administração Pública (SUAP), the institution's unified public administration system. To this end, a prototype was developed using Digital Image Processing techniques, Optical Character Recognition, regular expressions, document-based rules, user-assisted validation screens for verification and correction, data export in comma-separated values (CSV) format, and event logging. The evaluation was carried out in a controlled laboratory environment, without integration with SUAP and without operational validation by SRA staff, using 30 documents distributed among RG, CIN, and CNH, including documentary samples obtained through different means and synthetic documents. The results were classified as complete, partial, or failed, allowing the calculation of the Exact Field Match Rate and the Usable Recognition Rate. The Exact Field Match Rate, which considered only fully extracted fields, was 38.6% for RG, 67.5% for CIN, and 70.0% for CNH. As a complementary indicator, the Usable Recognition Rate, which included complete and partial results, was 75.7%, 78.8%, and 80.0%, respectively. Considering all document types, the overall Exact Field Match Rate was 60.4%, while the overall Usable Recognition Rate was 78.4%. Partial results still required user verification and manual correction. It is concluded that the proposed solution showed potential to support the initial extraction and organization of documentary information in a laboratory environment, but it still requires validation in the real SRA environment, user testing, expansion of the experimental dataset, and evaluation of possible secure integration with SUAP.

Keywords: digital image processing; optical character recognition; official documents; academic registration; digital transformation.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE SIGLAS

BGR	Blue, Green, Red (azul, verde e vermelho)
CEGE	Comitê Executivo de Governo Eletrônico
CIN	Carteira de Identidade Nacional
CNH	Carteira Nacional de Habilitação
CPF	Cadastro de Pessoas Físicas
CSV	Comma-Separated Values (valores separados por vírgula)
e-ARQ Brasil	Modelo de Requisitos para Sistemas Informatizados de Gestão Arquivística de Documentos
IF Baiano	Instituto Federal de Educação, Ciência e Tecnologia Baiano
JPEG	Joint Photographic Experts Group
LGPD	Lei Geral de Proteção de Dados Pessoais
OCR	Optical Character Recognition (Reconhecimento Óptico de Caracteres)
PDCA	Plan, Do, Check, Act (Planejar, Executar, Verificar e Agir)
PDI	Processamento Digital de Imagens
PDSII	Processo de Desenvolvimento de Software Iterativo e Incremental
PNG	Portable Network Graphics
PSM	Page Segmentation Mode (modo de segmentação de página)
RGB	Red, Green, Blue (vermelho, verde e azul)
RG	Registro Geral
SRA	Secretaria de Registros Acadêmicos
SUAP	Sistema Unificado de Administração Pública
TA	Taxa de Acerto
TRA	Taxa de Reconhecimento Aproveitável

SUMÁRIO

1 INTRODUÇÃO

1.1 Contextualização

Com o avanço das tecnologias e o aumento do uso de serviços digitais, a transformação digital passou a fazer parte do dia a dia das pessoas e das instituições. Nas organizações públicas, esse processo vem ocorrendo de forma gradual, nem sempre na velocidade necessária para atender às demandas atuais. Em muitos casos, existem atividades que dependem da conferência e da digitação manual de informações presentes em documentos oficiais.

No Instituto Federal de Educação, Ciência e Tecnologia Baiano (IF Baiano) – Campus Catu, durante os processos de matrícula, a Secretaria de Registros Acadêmicos (SRA) recebe documentos dos estudantes por diferentes meios: imagens no formato Joint Photographic Experts Group (JPEG), com extensões `.jpg` ou `.jpeg`, e no formato Portable Network Graphics (PNG), recebidas por e-mail, além de fotografias digitais e documentos apresentados presencialmente. Entre os documentos mais utilizados estão o Registro Geral (RG), a Carteira Nacional de Habilitação (CNH), certidões de nascimento e históricos escolares.

O RG, atualmente substituído gradativamente pela Carteira de Identidade Nacional (CIN), é um documento oficial destinado à identificação civil dos cidadãos brasileiros. Já a CNH, além de autorizar a condução de veículos automotores, possui validade como documento oficial de identificação em todo o território nacional. Esses documentos contêm informações para o cadastramento de estudantes: nome completo, número de inscrição no Cadastro de Pessoas Físicas (CPF), data de nascimento, filiação e demais dados pessoais utilizados em processos administrativos.

Atualmente, as informações contidas nesses documentos precisam ser consultadas e digitadas manualmente no Sistema Unificado de Administração Pública (SUAP)[1].

Segundo informações obtidas junto à SRA do IF Baiano – Campus Catu, o processo de matrícula exige o preenchimento de múltiplas telas do sistema,

[1] Plataforma de gestão acadêmica e administrativa utilizada pelos Institutos Federais e outras instituições públicas de ensino

demandando aproximadamente 7 (sete) minutos para o cadastramento de cada estudante. Além de tornar a atividade repetitiva e consumir tempo dos servidores responsáveis, erros durante o preenchimento podem gerar retrabalho, uma vez que determinadas informações precisam ser inseridas novamente quando ocorre alguma inconsistência ou interrupção do processo.

Diante dessa circunstância, surge a necessidade de um modelo computacional capaz de auxiliar a captura automática das informações presentes nos documentos utilizados durante o processo de matrícula.

A utilização de técnicas de Processamento Digital de Imagens (PDI) e Reconhecimento Óptico de Caracteres (OCR) pode contribuir para reduzir a quantidade de dados digitados manualmente, além de facilitar a conferência das informações antes de sua utilização nos sistemas institucionais.

Entretanto, a extração automática de dados não é uma tarefa simples, pois fatores relacionados à qualidade da imagem, iluminação, sombras, resolução e diferentes modelos de documentos podem interferir diretamente nos resultados obtidos. Dessa forma, além da identificação automática dos campos nos referidos documentos, torna-se importante disponibilizar mecanismos que permitam a conferência e a correção das informações extraídas.

Nesse sentido, este trabalho propõe o desenvolvimento de um modelo computacional para extrair automaticamente informações, especificamente, do RG, da CIN e da CNH.

1.2 Problema de Pesquisa

O processo de matrícula da SRA do IF Baiano – Campus Catu depende da conferência e da digitação manual de informações presentes em documentos oficiais. Essa atividade demanda tempo dos servidores responsáveis pelo cadastramento dos estudantes no processo de matrícula e está sujeita a retrabalho decorrente de erros de preenchimento.

Diante desse contexto, formula-se a seguinte questão de pesquisa: em que medida um modelo computacional baseado em Processamento Digital de Imagens, Reconhecimento Óptico de Caracteres, regras documentais e validação assistida pelo usuário pode auxiliar a captura, a conferência e a organização de informações presentes em RG, CIN e CNH no processo de matrícula da SRA do IF Baiano –

Campus Catu?

1.3 Hipótese

Parte-se da hipótese de que a integração entre técnicas de Processamento Digital de Imagens, Reconhecimento Óptico de Caracteres, regras documentais e validação assistida pelo usuário permite reconhecer e organizar, de forma integral ou parcialmente aproveitável, parte dos campos presentes em RG, CIN e CNH, oferecendo apoio à etapa inicial do cadastramento acadêmico, sem eliminar a necessidade de conferência humana.

1.4 Justificativa

O desenvolvimento deste trabalho justifica-se pela necessidade de apoiar atividades de cadastramento que dependem da consulta e da inserção manual de informações presentes em documentos oficiais. No processo de matrícula estudantil, os documentos RG, CIN e CNH são frequentemente utilizados para obtenção de dados pessoais e de identificação.

Durante os procedimentos de matrícula realizados pela SRA do IF Baiano – Campus Catu, as informações presentes nesses documentos precisam ser consultadas e transcritas manualmente para o SUAP, tornando a atividade repetitiva e suscetível a erros de preenchimento.

O cadastramento de um estudante, realizado pela SRA, pode demandar aproximadamente 7 (sete) minutos devido ao preenchimento de múltiplas telas do sistema, aumentando o tempo necessário para a execução da atividade e potencializando a ocorrência de retrabalho em situações de erro ou inconsistência dos dados inseridos.

Dessa forma, torna-se relevante desenvolver um modelo computacional que auxilie a captura e a conferência das informações presentes nesses documentos, reduzindo a quantidade de dados digitados manualmente e disponibilizando as informações extraídas para validação antes de sua utilização.

Espera-se que a solução proposta contribua para apoiar atividades relacionadas ao cadastramento de estudantes, além de servir de referência para futuras aplicações voltadas à SRA do IF Baiano – Campus Catu.

1.5 Objetivo Geral

Desenvolver um modelo computacional capaz de auxiliar o processo de matrícula e cadastramento de estudantes na SRA do IF Baiano — Campus Catu, mediante a identificação, a extração, a validação por regras documentais e a organização de informações presentes em RG, CIN e CNH, com conferência assistida pelo usuário.

1.6 Objetivos Específicos

Para atingir o objetivo geral desta pesquisa, foram definidos os seguintes objetivos específicos:

a) estudar técnicas de Processamento Digital de Imagens aplicáveis à extração automática de informações presentes em RG, CIN e CNH;
b) aplicar técnicas de pré-processamento para melhoria da qualidade das imagens utilizadas pelo OCR;
c) implementar um modelo de reconhecimento e extração automática de dados utilizando OCR;
d) investigar mecanismos de validação das informações extraídas para reduzir erros de reconhecimento;
e) avaliar o desempenho da solução em ambiente laboratorial controlado utilizando imagens de documentos oficiais brasileiros.

1.7 Organização do trabalho

Metodologicamente, esta pesquisa caracteriza-se como aplicada, exploratória e de caráter experimental. O estudo envolveu o desenvolvimento de um protótipo computacional, orientado pelo Processo de Desenvolvimento de Software Iterativo e Incremental (PDSII) e organizado com o apoio do ciclo Plan-Do-Check-Act (PDCA).

A avaliação foi realizada em ambiente laboratorial controlado, mediante testes com 30 documentos dos tipos RG, CIN e CNH, sem integração ao SUAP e sem validação operacional com os servidores da SRA.

Além desta introdução, o trabalho está organizado em seis capítulos. O Capítulo 2 apresenta a fundamentação teórica sobre transformação digital, gestão

documental, visão computacional, Processamento Digital de Imagens e Reconhecimento Óptico de Caracteres. O Capítulo 3 examina trabalhos correlatos. O Capítulo 4 descreve os procedimentos metodológicos adotados. O Capítulo 5 detalha o modelo computacional desenvolvido. O Capítulo 6 apresenta e discute os resultados experimentais. Por fim, o Capítulo 7 reúne as considerações finais, as limitações da pesquisa e as possibilidades de trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 Transformação digital na administração pública

A transformação digital tem se consolidado como um dos principais instrumentos de modernização das organizações públicas e privadas. Seu objetivo ultrapassa a simples informatização de procedimentos, envolvendo mudanças estruturais nos processos organizacionais, na gestão da informação e na prestação de serviços aos usuários.

No setor público, a transformação digital está associada à busca por maior eficiência administrativa, transparência, acessibilidade e melhoria da qualidade dos serviços oferecidos à sociedade. A adoção de tecnologias digitais possibilita a automatização de atividades repetitivas, a redução de custos operacionais e a otimização dos fluxos de trabalho.

Segundo Ávila, Lanza e Valotto (2023):

> A Transformação Digital no setor público brasileiro teve como marco inicial a criação do Comitê Executivo de Governo Eletrônico (CEGE), no ano de 2000, como uma das consequências do modelo gerencialista adotado pelo Governo Federal brasileiro em alinhamento às novas diretrizes trazidos pela Emenda Constitucional 19/98 (BRASIL, 1998), que incorporou o princípio da eficiência da administração pública e outros comandos voltados ao aprimoramento da máquina estatal (Ávila; Lanza; Valotto, 2023, p. 60).

Segundo Schallmo et al. (2021), a transformação digital não deve ser compreendida apenas como a adoção de novas tecnologias, mas como um processo de reestruturação organizacional capaz de gerar novas formas de produção, compartilhamento e utilização da informação.

> A transformação digital impacta nas mais diferentes áreas da sociedade e da economia. Ela abre novas oportunidades de networking e possibilita a colaboração entre diversos atores, que podem, assim, trocar informações e, consequentemente, iniciar novos processos (Schallmo et al., 2021, p. 15).

Os autores ainda esclarecem que:

> A grande diferença com a transformação digital é como os funcionários interpretam o know-how recém-adquirido e o usam para tomar decisões melhores. Todas as novas fontes de dados podem gerar conhecimento novo baseado nesses dados. Em vez de apenas tornar os processos mais eficientes ou rápidos, que é o objetivo da automação, a transformação digital

> exige que os indivíduos repensem processos antigos e reinventem novos processos e decisões (Schallmo et al., 2021, p. 22).

Escobar (2020) destaca que a digitalização dos serviços públicos contribui para ampliar a eficiência administrativa, melhorar a experiência dos usuários e aumentar a capacidade das instituições em responder às demandas da sociedade contemporânea.

> Em especial no setor público, a transformação digital significa novas maneiras de trabalhar com as partes interessadas, construindo novas estruturas de prestação de serviços e criando novas formas de relacionamento (Escobar, 2020, p. 96).

De forma complementar, Caserta (2023) ressalta que a transformação digital depende da integração entre tecnologia, pessoas e processos, exigindo mudanças culturais e organizacionais que favoreçam a inovação e a utilização estratégica dos dados.

> Ela acontece quando as organizações e as pessoas passam a enxergar a necessidade de fazer parte de um alinhamento de serviços ou de produtos por meio da tecnologia. Ou seja, a transformação digital não se trata do desenvolvimento de alguma tecnologia em si; ela acontece quando determinada tecnologia é usada como meio para que uma evolução aconteça.
> [...]
> Transformação digital é basicamente isso. Para as pessoas, há a mudança no hábito e, para as empresas, é o processo pelo qual elas precisam passar para que consigam entregar produtos ou serviços de uma maneira que esteja alinhada com a sociedade atual – uma sociedade totalmente impactada pela revolução tecnológica (Caserta, 2023, p. 31-32).

Caserta (2023) ainda complementa:

> Acredito que a partir de agora a tecnologia precisa entrar como um dos pilares estratégicos para a transformação digital, ainda que ela sozinha não seja responsável por toda a transformação. Pois, sem ela, você não encontrará meios de escalar e otimizar atividades em sua organização.
> [...]
> Considerar a tecnologia como pilar estratégico é fundamental, porque é ela que permite a criação de novos produtos e inovações de uma maneira muito rápida. Também é ela que permite que os conceitos sejam testados e validados com a mesma rapidez.
> Sem a tecnologia, dificilmente você conseguirá fazer com que todos os demais pilares e objetivos tenham uma evolução em potencial (Caserta, 2023, p. 150-151).

No contexto das instituições públicas de ensino, a transformação digital tem

possibilitado a modernização de processos acadêmicos e administrativos, incluindo matrículas, gestão documental, atendimento aos estudantes e controle acadêmico. Entretanto, esse processo não se limita à simples digitalização de documentos ou à informatização de rotinas existentes, envolvendo a utilização estratégica das tecnologias digitais para otimizar processos, melhorar a qualidade dos serviços prestados e ampliar a eficiência institucional.

Apesar desses avanços, diversas atividades ainda dependem da conferência manual de documentos e da digitação de informações em sistemas institucionais. Nesse cenário, tecnologias de Visão Computacional, Processamento Digital de Imagens e Reconhecimento Óptico de Caracteres podem contribuir para automatizar tarefas documentais, reduzir o retrabalho, diminuir o tempo de processamento das informações e aumentar a eficiência administrativa.

2.2 Gestão documental e automação de processos acadêmicos

A gestão documental não se limita ao armazenamento de arquivos ou à conversão de documentos físicos para o formato digital. Nos termos do art. 3º da Lei nº 8.159/1991:

> Art. 3º - Considera-se gestão de documentos o conjunto de procedimentos e operações técnicas referentes à sua produção, tramitação, uso, avaliação e arquivamento em fase corrente e intermediária, visando a sua eliminação ou recolhimento para guarda permanente. (Brasil, 1991, art. 3º).

Essa concepção considera o documento durante todo o seu ciclo de vida e relaciona sua organização às atividades institucionais que o produziram ou receberam.

Em ambientes informatizados, a gestão arquivística deve contribuir para que os documentos sejam produzidos e mantidos de forma confiável e autêntica, permanecendo acessíveis pelo tempo necessário. O Modelo de Requisitos para Sistemas Informatizados de Gestão Arquivística de Documentos (e-ARQ Brasil) constitui uma especificação voltada às organizações produtoras ou recebedoras de documentos, aos sistemas de gestão arquivística e aos próprios documentos, com o objetivo de assegurar essas características (Conselho Nacional de Arquivos, 2022, p. 10-11).

O modelo aplica-se a sistemas que compreendem documentos digitais, não

digitais e híbridos. A gestão arquivística abrange o ciclo de vida documental, incluindo a produção e o recebimento, a tramitação, o uso, a avaliação, o arquivamento e a destinação final. Além dessas operações, o e-ARQ Brasil estabelece requisitos e procedimentos relacionados à captura, ao controle de acesso, às trilhas de auditoria, à segurança e à preservação dos documentos (Conselho Nacional de Arquivos, 2022, p. 11, 23-24, 34, 41-43, 97-98).

A implantação desses procedimentos não depende exclusivamente da adoção de um software. O próprio e-ARQ Brasil destaca que sua implementação requer a interação entre profissionais das áreas de administração, arquivo e tecnologia da informação, considerando o contexto e as necessidades específicas de cada instituição (Conselho Nacional de Arquivos, 2022, p. 12).

Nos processos acadêmicos, os documentos apresentados pelos estudantes integram fluxos administrativos destinados a comprovar informações e subsidiar a formação dos registros institucionais. Durante a matrícula, dados provenientes de documentos de identificação são conferidos e transcritos para o sistema acadêmico.

Dessa forma, uma falha na captura ou na interpretação dessas informações pode ser propagada para o cadastro institucional, gerando inconsistências e necessidade de correção posterior. A utilização de ferramentas computacionais nesse contexto deve, portanto, preservar a possibilidade de conferência do documento original, a identificação da origem dos dados e a responsabilidade do agente encarregado de validar as informações.

A automação administrativa consiste na aplicação de recursos tecnológicos para executar ou apoiar etapas de um processo que anteriormente dependiam predominantemente de intervenção manual. No setor público, a Lei nº 14.129/2021 relaciona o Governo Digital à desburocratização, à modernização, à simplificação dos serviços e ao aumento da eficiência administrativa.

> Art. 1º Esta Lei dispõe sobre princípios, regras e instrumentos para o aumento da eficiência da administração pública, especialmente por meio da desburocratização, da inovação, da transformação digital e da participação do cidadão. (Brasil, 2021, art. 1º).

Entretanto, a automação de uma atividade isolada não representa, por si só, a transformação integral do processo. Para produzir resultados institucionais, a tecnologia precisa ser incorporada aos procedimentos, às responsabilidades e aos

mecanismos de controle existentes.

Também é necessário distinguir a digitalização documental do reconhecimento automatizado de conteúdo. A digitalização produz uma representação digital de um documento físico, enquanto o OCR busca converter os caracteres visualmente presentes na imagem em dados textuais que possam ser processados pelo sistema.

Ademais, o Decreto nº 10.278/2020 estabelece “a técnica e os requisitos para a digitalização de documentos públicos ou privados, a fim de que os documentos digitalizados produzam os mesmos efeitos legais dos documentos originais”. O referido diploma legal determina ainda as regras gerais de digitalização:

> Art. 4º Os procedimentos e as tecnologias utilizados na digitalização de documentos físicos devem assegurar:
> I - a integridade e a confiabilidade do documento digitalizado;
> II - a rastreabilidade e a auditabilidade dos procedimentos empregados;
> III - o emprego dos padrões técnicos de digitalização para garantir a qualidade da imagem, da legibilidade e do uso do documento digitalizado;
> IV - a confidencialidade, quando aplicável; e
> V - a interoperabilidade entre sistemas informatizados. (Brasil, 2020, art. 4º).

Por sua vez, o resultado produzido pelo OCR não substitui automaticamente o documento original nem assegura, isoladamente, a correção do conteúdo reconhecido, razão pela qual deve ser submetido à conferência e à validação.

Os documentos analisados nesta pesquisa contêm dados pessoais, como nome, CPF, data de nascimento, filiação e número de identificação. Por esse motivo, seu processamento deve observar princípios de finalidade, adequação, necessidade, segurança e prevenção previstos na Lei Geral de Proteção de Dados Pessoais (LGPD), Lei nº 13.709/2018 (Brasil, 2018, art. 6º).

No desenvolvimento de soluções documentais, esses princípios se relacionam à limitação dos dados processados, à restrição de acesso, à proteção contra usos não autorizados e à manutenção dos arquivos somente pelo período necessário à finalidade definida.

No contexto desta pesquisa, o modelo computacional desenvolvido não constitui um Sistema Informatizado de Gestão Arquivística de Documentos e não foi integrado ao SUAP. Sua finalidade está limitada ao apoio à etapa inicial de captura, identificação, validação e organização de dados presentes em RG, CIN e CNH. O documento original permanece como fonte de referência, e o usuário continua responsável pela conferência e pela confirmação das informações antes de seu

armazenamento. Assim, a solução deve ser compreendida como uma ferramenta de apoio inserida em um processo documental e administrativo mais amplo, e não como substituta da gestão documental institucional ou da atuação dos servidores.

2.3 Visão computacional

A Visão Computacional é uma área da Ciência da Computação que busca desenvolver métodos capazes de permitir que sistemas computacionais interpretem e compreendam informações presentes em imagens e vídeos digitais.

Seu desenvolvimento resulta da integração entre diferentes áreas do conhecimento, incluindo processamento digital de imagens, inteligência artificial, reconhecimento de padrões e aprendizado de máquina. O objetivo principal consiste em reproduzir computacionalmente parte da capacidade humana de percepção visual. Segundo Barelli (2018), a visão computacional complementa a visão biológica, ou seja, “a Visão Computacional estuda e implementa sistemas capazes de enxergar por meio de processos artificiais, implementados por hardwares e softwares” (Barelli, 2018, p. 3).

Atualmente, aplicações de Visão Computacional estão presentes em sistemas de reconhecimento facial, inspeção industrial, diagnóstico médico por imagem, monitoramento de tráfego, biometria e reconhecimento documental.

No contexto desta pesquisa, a Visão Computacional fornece a base tecnológica necessária para o tratamento das imagens dos documentos oficiais utilizados durante os processos de matrícula, possibilitando a identificação e a extração automatizada das informações presentes nesses documentos.

2.4 Processamento digital de imagens (PDI)

O Processamento Digital de Imagens (PDI) corresponde ao conjunto de técnicas computacionais destinadas à manipulação, melhoria e análise de imagens digitais. Seu principal objetivo consiste em tornar as imagens mais adequadas à interpretação humana ou ao processamento automatizado por sistemas computacionais.

Para Albuquerque e Albuquerque (2002), o objetivo do processamento da imagem é extrair as informações existentes nela:

> A disciplina ‘Processamento de Imagens’ vem na realidade do Processamento de Sinais. Os sinais, como as imagens, são na realidades [sic] um suporte físico que carrega no seu interior uma determinada INFORMAÇÃO. Esta INFORMAÇÃO pode estar associada a uma medida (neste caso falamos de um sinal em associação a um fenômeno físico), ou pode estar associada à um nível cognitivo (neste caso falamos de conhecimento). Processar uma imagem consiste em transformá-la sucessivamente com o objetivo de extrair mais facilmente a INFORMAÇÃO nela presente (Albuquerque; Albuquerque, 2002, p. 1).

Segundo Queiroz e Gomes (2001):

> O Processamento Digital de Imagens (PDI) não é uma tarefa simples, na realidade envolve um conjunto de tarefas interconectadas. Tudo se inicia com a captura de uma imagem, a qual, normalmente, corresponde à iluminação que é refletida na superfície dos objetos, realizada através e [sic] um sistema de aquisição. Após a captura por um processo de digitalização, uma imagem precisa ser representada de forma apropriada para tratamento computacional (Queiroz; Gomes, 2001, p. 2).

Para os autores, o processamento digital de imagens procura adequar as imagens originais para o processamento computacional. “Na área de processamento digital de imagens existem muitos problemas que necessitam de técnicas capazes de melhorar a qualidade das imagens para que possam ser feitas análises e interpretações por profissionais” (Queiroz; Gomes, 2001, p. 30).

No contexto desta pesquisa, o Processamento Digital de Imagens desempenha papel fundamental na preparação das imagens dos documentos submetidos ao OCR, contribuindo para melhorar a legibilidade dos caracteres e aumentar a precisão da extração textual.

2.4.1 Conversão para tons de cinza

Uma das primeiras etapas do pré-processamento consiste na conversão das imagens coloridas para escala de cinza. Essa transformação reduz a quantidade de informações que precisam ser processadas, simplificando as etapas posteriores de análise.

Segundo Marques Filho e Vieira Neto (1999), as informações de cor representam um importante elemento na análise automática de imagens, podendo contribuir para a identificação e segmentação de objetos.

> O uso de cores em processamento digital de imagens decorre de dois fatores motivantes principais:
> 1. Na análise automática de imagens (reconhecimento de padrões), a cor é um poderoso descritor das propriedades de um objeto, que pode simplificar sua identificação e segmentação (Marques Filho; Vieira Neto, 1999, p. 118).

Barelli (2018) afirma que as “imagens em tons de cinza são consideradas uma boa alternativa, pois torna o processamento mais rápido, uma vez que há menos informações para serem processadas” (Barelli, 2018, p. 40).

Ainda segundo Barelli (2018):

> As imagens coloridas, diferente da em tons de cinza, utilizam de mais de um canal para serem representadas. O espaço de cor RGB é um dos modelos para representação de imagens coloridas, e ele utiliza três canais diferentes, cada um para representar a intensidade de uma cor primária em cada pixel da imagem (Barelli, 2018, p. 53).

Na sigla RGB, as letras correspondem aos termos ingleses *red*, *green* e *blue*, isto é, vermelho, verde e azul.

A conversão para tons de cinza é particularmente relevante, considerando que os documentos de identificação brasileiros, incluindo a CNH, possuem elementos visuais e textuais que mantêm contraste suficiente mesmo após essa conversão.

No contexto desta pesquisa, essa etapa contribui para a redução de informações cromáticas desnecessárias, permitindo destacar os caracteres textuais e melhorar a qualidade das imagens submetidas ao OCR, favorecendo maior precisão na extração automatizada dos dados presentes nos documentos analisados.

2.4.2 Segmentação por binarização

A segmentação por binarização é uma técnica amplamente utilizada para separar os elementos de interesse do fundo da imagem. Em documentos oficiais, essa etapa favorece a identificação dos caracteres textuais e reduz interferências visuais.

> A segmentação por binarização resulta em uma imagem binária, na qual geralmente o objeto de interesse é representado pela cor branca e o segundo plano pela preta. Esse procedimento de binarização, ou aplicação de limiar de intensidade, pode ser realizado pela função threshold da biblioteca OpenCV (Barelli, 2018, p. 160-161).

Gonzalez e Woods (2018, p. 763) afirmam que "escolher a média global geralmente fornece melhores resultados quando o fundo é quase constante e todas

as intensidades do objeto estão acima ou abaixo da intensidade do fundo", e continuam explicando que “esta implementação é útil em aplicações de processamento de documentos, onde a velocidade é um requisito fundamental”.

Para esta pesquisa foi adotado o método de Otsu, pois, segundo Barelli (2018, p. 170), “esse algoritmo pode ser usado junto à função threshold, necessitando apenas que a constante THRESH_OTSU seja somada ao tipo definido para a binarização”.

2.4.3 Filtragem de ruídos

A filtragem de ruídos busca eliminar interferências presentes na imagem que possam comprometer o reconhecimento dos caracteres.

Segundo Barelli (2018), “o filtro de mediana é não linear e apresenta bons resultados no tratamento de ruído do tipo ‘sal e pimenta’" (Barelli, 2018, p. 117).

O autor complementa que “matematicamente, este filtro atua alterando o valor de cada pixel-alvo pela mediana estatística dos valores dos pixels vizinhos” e que “o método medianBlur da biblioteca OpenCV nos permite aplicá-lo a imagens” (Barelli, 2018, p. 118).

Ainda segundo Barelli (2018) “o primeiro deles é a imagem que receberá o tratamento, e o segundo é um valor inteiro, ímpar e positivo, que indicará a sua intensidade” (Barelli, 2018, p. 118).

2.4.4 Aplicações do PDI em documentos oficiais

O Processamento Digital de Imagens possui ampla aplicação em sistemas de digitalização documental, reconhecimento de padrões, leitura automática de formulários e gestão eletrônica de documentos. Em documentos oficiais, técnicas de pré-processamento são utilizadas para corrigir imperfeições decorrentes da captura das imagens, melhorar a legibilidade dos caracteres e aumentar a eficiência dos sistemas de reconhecimento automático.

Problemas relacionados a sombras, reflexos, inclinação do documento, baixa resolução e variações de iluminação podem comprometer significativamente o desempenho dos mecanismos de reconhecimento textual. Por esse motivo, etapas de conversão para tons de cinza, binarização e filtragem de ruídos são frequentemente empregadas para preparar as imagens antes da execução do OCR.

No contexto desta pesquisa, essas técnicas foram aplicadas em documentos de identificação brasileiros: o RG, a CIN e a CNH, com o objetivo de aumentar a precisão da extração automática das informações presentes nesses documentos.

2.5 Reconhecimento óptico de caracteres (OCR)

2.5.1 Conceitos fundamentais

O Reconhecimento Óptico de Caracteres (Optical Character Recognition – OCR) consiste em uma tecnologia capaz de converter imagens contendo texto em informações textuais manipuláveis por sistemas computacionais.

Essa tecnologia permite transformar documentos digitalizados, fotografias e imagens em dados estruturados, possibilitando sua edição, armazenamento, pesquisa e processamento automatizado.

2.5.2 Funcionamento do OCR

O funcionamento do OCR envolve diversas etapas sucessivas. Inicialmente, a imagem passa por procedimentos de pré-processamento destinados a melhorar sua qualidade visual. Em seguida, ocorre a identificação das regiões contendo texto, a segmentação dos caracteres e o reconhecimento dos padrões visuais encontrados.

Segundo Smith (2007), o Tesseract OCR utiliza um pipeline de processamento composto por múltiplas etapas. Inicialmente, realiza-se uma análise dos componentes conectados da imagem, permitindo identificar estruturas visuais associadas aos caracteres. Posteriormente, o mecanismo organiza essas informações em regiões textuais e realiza o reconhecimento dos caracteres, utilizando estratégias adaptativas para aprimorar a identificação do texto.

> O processamento segue um pipeline tradicional passo a passo, mas algumas das etapas eram incomuns na época e possivelmente ainda são hoje. O primeiro passo é uma análise de componentes conectados, na qual os contornos dos componentes são armazenados. Essa foi uma decisão de projeto computacionalmente dispendiosa na época, mas tinha uma vantagem significativa: pela inspeção do aninhamento dos contornos e do número de contornos de filhos e netos, é simples detectar texto invertido e reconhecê-lo tão facilmente quanto texto preto sobre fundo branco. [...].
> O reconhecimento então prossegue como um processo de duas etapas. Na primeira etapa, tenta-se reconhecer cada palavra por vez. Cada palavra que for considerada satisfatória é passada para um classificador adaptativo como

> dados de treinamento [...]. (Smith, 2007, p. 1).

Smith (2007) descreve a arquitetura histórica do Tesseract, baseada em análise de componentes conectados e reconhecimento adaptativo. Posteriormente, a partir da versão 4, o mecanismo passou a incorporar um motor de reconhecimento baseado em redes neurais de memória de longo e curto prazo (*Long Short-Term Memory*). A versão 5 utilizada nesta pesquisa mantém essa arquitetura neural como uma das opções de reconhecimento (Tesseract OCR, [s. d.]). Assim, o estudo de Smith é empregado como referência histórica, enquanto a descrição da versão utilizada deve considerar também a documentação atual do Tesseract.

Após o reconhecimento dos caracteres, o conteúdo textual é convertido em formato digital, possibilitando seu armazenamento, validação e utilização por outros sistemas computacionais.

2.5.3 Aplicações do OCR em documentos

O Reconhecimento Óptico de Caracteres é amplamente utilizado em processos de digitalização e automação documental. Suas aplicações incluem leitura automática de formulários, processamento de contratos, digitalização de arquivos históricos, reconhecimento de notas fiscais, sistemas bancários, protocolos eletrônicos e identificação automática de documentos pessoais.

Na administração pública, o OCR vem sendo empregado para reduzir a necessidade de digitação manual, aumentar a produtividade dos servidores e acelerar atividades relacionadas à conferência e cadastramento de informações.

No contexto das instituições públicas de ensino, o OCR pode ser utilizado na solução de apoio aos setores responsáveis pelo cadastramento acadêmico, automatizando a captura de informações presentes em documentos de identificação dos estudantes: RG, CIN e CNH. Dessa forma, reduz-se a necessidade de transcrição manual desses dados para sistemas institucionais, diminuindo o tempo de processamento e a possibilidade de erros de digitação.

2.5.4 Limitações do OCR

Apesar dos avanços observados nos mecanismos modernos de reconhecimento óptico de caracteres, o desempenho do OCR permanece fortemente

dependente da qualidade das imagens processadas.

Fatores de baixa resolução, iluminação inadequada, sombras, reflexos, inclinação do documento, compressão excessiva da imagem e degradação dos caracteres podem reduzir significativamente a precisão do reconhecimento textual.

Além das características da imagem, diferenças de layout entre documentos, variações tipográficas e a presença de elementos gráficos complexos também podem dificultar a identificação correta das informações.

Em razão dessas limitações, sistemas baseados em OCR geralmente utilizam técnicas complementares de Processamento Digital de Imagens, validação de dados e conferência humana para aumentar a confiabilidade dos resultados produzidos.

2.5.5 O Tesseract OCR

O Tesseract OCR é um mecanismo de reconhecimento óptico de caracteres de código aberto originalmente desenvolvido nos laboratórios da Hewlett-Packard entre 1984 e 1994.

Atualmente, o Tesseract é considerado uma das ferramentas OCR mais utilizadas em projetos acadêmicos e profissionais devido à sua robustez, flexibilidade e suporte a múltiplos idiomas.

Nesta pesquisa, o Tesseract foi empregado na função de principal mecanismo de reconhecimento textual, sendo responsável pela extração automatizada das informações presentes nos documentos analisados. Seu desempenho mostrou-se diretamente relacionado à qualidade das imagens submetidas ao processamento, reforçando a importância das etapas de Processamento Digital de Imagens realizadas previamente.

3 TRABALHOS CORRELATOS

Esta seção apresenta estudos relacionados à utilização de Processamento Digital de Imagens e Reconhecimento Óptico de Caracteres aplicados à extração automatizada de informações textuais. Os trabalhos selecionados abordam diferentes estratégias de automação documental, reconhecimento de caracteres e validação de dados em imagens digitais.

3.1 Trabalho de Leite (2024)

Embora o trabalho de Leite (2024) tenha sido desenvolvido em contexto distinto do desta pesquisa, o estudo foi considerado relevante por empregar técnicas de Processamento Digital de Imagens semelhantes às utilizadas neste trabalho.

O autor apresenta um método para identificação e segmentação de objetos em imagens utilizando etapas de pré-processamento, detecção de bordas e operações morfológicas.

A relação com esta pesquisa não está no objeto analisado, mas nas técnicas empregadas durante o processamento das imagens. Leite (2024) utiliza etapas de pré-processamento, detecção de bordas e operações morfológicas para identificar e segmentar regiões de interesse.

De maneira semelhante, neste trabalho são aplicadas técnicas de conversão para tons de cinza, binarização, redução de ruídos e segmentação de regiões contendo informações textuais antes da execução do OCR.

Em ambos os casos, a qualidade das etapas de tratamento da imagem influencia diretamente os resultados obtidos nas fases seguintes do processamento. No trabalho de Leite (2024), essas etapas contribuem para a identificação e a segmentação dos objetos presentes na imagem. Já na presente pesquisa, o pré-processamento busca melhorar a legibilidade dos documentos antes da aplicação do OCR, favorecendo a extração de informações como nome, CPF, número do documento e data de nascimento.

Embora o objeto analisado por Leite (2024) seja distinto, o estudo contribui metodologicamente ao evidenciar a importância do pré-processamento e da segmentação das regiões de interesse antes da etapa principal de análise computacional.

3.2 Trabalho de Barbosa (2017)

Barbosa (2017) apresentou um sistema capaz de realizar o reconhecimento automático de placas veiculares utilizando técnicas de Processamento Digital de Imagens e Reconhecimento Óptico de Caracteres (OCR). A proposta buscou automatizar a identificação de veículos a partir de imagens digitais, empregando métodos de tratamento de imagens, segmentação da região da placa e reconhecimento dos caracteres alfanuméricos presentes nela.

O trabalho empregou as seguintes técnicas:

a) conversão da imagem para tons de cinza utilizando a biblioteca OpenCV;
b) aplicação de limiarização (threshold) para binarização da imagem;
c) utilização do filtro Gaussian Blur para redução de ruídos e suavização da imagem;
d) segmentação da região de interesse por meio da detecção de contornos utilizando a função findContours;
e) identificação de regiões retangulares compatíveis com o formato de placas veiculares;
f) recorte da região segmentada utilizando técnicas de Região de Interesse;
g) aplicação do mecanismo Tesseract OCR para reconhecimento dos caracteres presentes na placa;
h) avaliação quantitativa dos resultados obtidos por meio da comparação entre o texto reconhecido e o texto real presente nas placas analisadas.

Os resultados obtidos demonstraram uma taxa média de acerto de aproximadamente 77,5% no reconhecimento das placas veiculares analisadas. O autor observou que fatores como qualidade da imagem, luminosidade, presença de ruídos e características da fonte utilizada influenciaram diretamente o desempenho do OCR.

A correlação entre o trabalho de Barbosa (2017) e a presente pesquisa está na utilização conjunta de técnicas de Processamento Digital de Imagens e do mecanismo Tesseract OCR para extração automática de informações textuais a partir de imagens.

Assim como nesta pesquisa, Barbosa (2017) realiza etapas de pré-processamento das imagens antes da aplicação do OCR, empregando técnicas de conversão para tons de cinza, binarização, redução de ruídos e segmentação de regiões de interesse. Em ambos os trabalhos, essas etapas têm a finalidade de

melhorar a qualidade visual da imagem e aumentar a precisão do reconhecimento textual.

Entretanto, existem diferenças importantes entre as propostas. Enquanto Barbosa (2017) utiliza o OCR para reconhecimento de caracteres presentes em placas veiculares, a presente pesquisa aplica o mesmo princípio para extração de informações em documentos oficiais brasileiros: RG, CIN e CNH.

Além disso, esta pesquisa incorpora etapas adicionais de identificação e validação de campos específicos, utilizando expressões regulares e regras práticas para localizar campos específicos conforme a estrutura de cada documento, incluindo informações: nome, CPF, número do documento, datas, sexo, naturalidade, órgão expedidor, categoria, nome do pai, nome da mãe e número de registro, conforme o tipo documental analisado, quando presente. Dessa forma, o OCR não representa o resultado final do processo, mas uma etapa intermediária dentro de um fluxo mais amplo de extração e validação automática de informações documentais.

3.3 Trabalho de Carvalho (2015)

O trabalho de Carvalho (2015) apresenta correlação direta com a presente pesquisa por utilizar uma arquitetura baseada em PDI e OCR para extração de informações textuais a partir de imagens digitais.

Durante o desenvolvimento da solução, o autor emprega técnicas de pré-processamento: limiarização (threshold), equalização de histograma, erosão, dilatação, abertura morfológica e fechamento morfológico, com o objetivo de melhorar a qualidade da imagem antes da etapa de reconhecimento textual. Essas técnicas desempenham função semelhante às utilizadas nesta pesquisa, que emprega conversão para tons de cinza, binarização, redução de ruídos e segmentação das regiões contendo informações textuais dos documentos.

Outra aproximação entre os trabalhos está na utilização do mecanismo Tesseract OCR para conversão das informações presentes na imagem em texto digital. Em ambos os casos, o OCR representa a etapa responsável por transformar os caracteres visuais em informações passíveis de processamento computacional.

Além do reconhecimento textual, a presente pesquisa incorpora etapas adicionais de identificação e validação automática dos dados extraídos. Após a execução do OCR, são aplicadas regras de validação baseadas em expressões

regulares e padrões documentais para localizar campos específicos conforme a estrutura de cada documento, incluindo informações: nome, CPF, número do documento, datas, sexo, naturalidade, órgão expedidor, categoria, nome do pai, nome da mãe e número de registro, conforme o tipo documental analisado, quando presente, etapa que não é contemplada na solução proposta por Carvalho (2015).

Entretanto, existem diferenças importantes entre as propostas. Carvalho (2015) utiliza técnicas de detecção de objetos, por meio de classificadores Haar Cascade, para localizar placas veiculares e realizar o reconhecimento dos caracteres presentes nelas. Já nesta pesquisa, o foco está na extração de informações presentes em documentos oficiais brasileiros: RG, CIN e CNH.

Dessa forma, embora os trabalhos sejam aplicados em contextos distintos, ambos compartilham técnicas semelhantes de processamento digital de imagens e reconhecimento textual, diferindo principalmente no tipo de objeto analisado e nos mecanismos de validação implementados após a etapa de OCR.

3.4 Trabalho de Uezima et al. (2018)

Uezima et al. (2018) tiveram como objetivo avaliar a utilização do mecanismo Tesseract OCR no reconhecimento de textos manuscritos em língua portuguesa. Para isso, os autores realizaram o treinamento do OCR utilizando amostras de escrita manual e posteriormente avaliaram seu desempenho por meio de métricas de acurácia e confidência.

A acurácia foi calculada utilizando a distância de Levenshtein, permitindo medir a similaridade entre o texto reconhecido pelo OCR e o texto original. Já a métrica de confidência foi utilizada para avaliar o grau de certeza atribuído pelo Tesseract às palavras reconhecidas durante o processamento.

O trabalho empregou as seguintes técnicas:

a) utilização do mecanismo Tesseract OCR para reconhecimento de caracteres;
b) treinamento do OCR com amostras de escrita manuscrita;
c) processamento de imagens contendo textos escritos manualmente;
d) avaliação quantitativa dos resultados utilizando a distância de Levenshtein para cálculo da acurácia;
e) análise dos níveis de confidência fornecidos pelo próprio Tesseract para

cada palavra reconhecida;

f) comparação do desempenho do OCR em textos escritos em letra de forma e textos cursivos.

Os resultados demonstraram melhor desempenho para textos escritos em letra de forma, enquanto textos cursivos apresentaram maiores dificuldades devido à conexão entre caracteres, dificultando sua segmentação e reconhecimento.

A correlação entre o trabalho de Uezima et al. (2018) e a presente pesquisa está na utilização do Tesseract OCR como mecanismo principal de reconhecimento textual a partir de imagens.

Enquanto Uezima et al. utilizam o OCR para reconhecer textos manuscritos, esta pesquisa emprega o mesmo mecanismo para reconhecer informações presentes em documentos oficiais brasileiros: RG, CIN e CNH.

Além da utilização do Tesseract OCR, ambos os trabalhos dependem diretamente da qualidade da imagem utilizada como entrada, uma vez que fatores como ruídos, contraste, resolução e legibilidade influenciam o desempenho do reconhecimento óptico de caracteres.

Entretanto, existem diferenças significativas entre as propostas. Uezima et al. concentram-se no treinamento e na avaliação do OCR para reconhecimento de escrita manual, utilizando métricas de acurácia baseadas na distância de Levenshtein e índices de confiança do próprio mecanismo OCR.

Já nesta pesquisa, o OCR é utilizado como uma etapa intermediária de um processo mais amplo, envolvendo pré-processamento das imagens, aplicação de múltiplos modos de segmentação de página (PSM), extração automática de campos específicos e validação dos dados reconhecidos por meio de expressões regulares e padrões documentais. Além disso, a avaliação da solução foi realizada pela comparação entre as informações presentes nos documentos originais e os dados extraídos pelo sistema, classificando os resultados como completos, parciais ou falhos, permitindo o cálculo de indicadores próprios de desempenho, denominados Taxa de Acerto e Taxa de Reconhecimento Aproveitável, conforme explicado na seção de metodologia.

3.5 Análise Comparativa

Os trabalhos analisados demonstram que o Processamento Digital de Imagens

é frequentemente utilizado como etapa preparatória para aumentar a qualidade das informações obtidas durante o processamento computacional.

Em Leite (2024), as técnicas de pré-processamento, detecção de bordas e operações morfológicas são utilizadas para segmentação de objetos. Já Barbosa (2017) e Carvalho (2015) empregam técnicas semelhantes para localizar e destacar regiões contendo caracteres antes da aplicação do OCR.

Observa-se também que Barbosa (2017), Carvalho (2015) e Uezima et al. (2018) utilizam o Tesseract OCR como mecanismo principal de reconhecimento textual.

Entretanto, os objetivos dos trabalhos são distintos. Enquanto Barbosa (2017) e Carvalho (2015) concentram-se na leitura de placas veiculares e Uezima et al. (2018) investigam o reconhecimento de textos manuscritos, a presente pesquisa utiliza o OCR para extrair informações contidas em documentos oficiais brasileiros.

Outro aspecto relevante é que nenhum dos trabalhos analisados realiza a identificação e validação de campos documentais específicos após a etapa de OCR. Em Barbosa (2017) e Carvalho (2015), o resultado final corresponde ao texto reconhecido na placa.

Em Uezima et al. (2018), a avaliação concentra-se na acurácia do reconhecimento de palavras manuscritas. Já nesta pesquisa, após a obtenção do texto pelo OCR, são aplicadas regras de validação baseadas em expressões regulares e padrões documentais para localizar informações do tipo: nome, CPF, número do documento, datas, sexo, naturalidade, órgão expedidor, categoria, nome do pai, nome da mãe e número de registro, conforme o tipo documental analisado, quando presente.

O Quadro 1 sistematiza a análise comparativa entre os trabalhos correlatos e esta pesquisa, considerando aspectos como o contexto de aplicação, as técnicas de PDI empregadas, a utilização do mecanismo OCR, as estratégias de extração das informações e os métodos de validação dos resultados.

A comparação evidencia que, embora os trabalhos anteriores utilizem técnicas semelhantes de PDI e, em alguns casos, o Tesseract OCR como ferramenta de reconhecimento textual, suas aplicações concentram-se na identificação de objetos, leitura de placas veiculares ou reconhecimento de textos manuscritos.

Diferentemente dessas abordagens, a presente pesquisa propõe um fluxo de processamento aplicado a documentos oficiais brasileiros, integrando pré-

processamento das imagens, reconhecimento textual por OCR, identificação automática de campos específicos e validação das informações extraídas por meio de expressões regulares e padrões documentais.

Quadro 1 - Análise comparativa entre os trabalhos correlatos e esta pesquisa

Autor	Trabalho	Técnicas de processamento digital	OCR	Estratégias de extração	Estratégia de validação	Contribuição para esta pesquisa
Leite (2024)	Implementação de algoritmo de processamento digital de imagens para segmentação de ovos de galinha.	Pré-processamento, detecção de bordas, operações morfológicas e segmentação de regiões de interesse.	Não utiliza OCR.	Extração baseada na identificação de objetos.	Não realiza validação textual.	Demonstrou a importância das etapas de segmentação e pré-processamento de imagens.
Barbosa (2017)	Processamento Digital de Imagens para o reconhecimento de placas de veículos.	Conversão para tons de cinza, binarização, filtro Gaussian Blur, segmentação por contornos e definição de ROI.	Tesseract OCR.	Extração dos caracteres presentes nas placas.	Comparação entre texto reconhecido e texto real.	Evidenciou a utilização integrada de OpenCV e Tesseract OCR para reconhecimento textual.
Carvalho (2015)	Identificação óptica de caracteres de placas de trânsito.	Equalização de histograma, limiarização, erosão, dilatação, abertura e fechamento morfológico, Haar Cascade.	Tesseract OCR.	Reconhecimento dos caracteres da placa após localização automática da região de interesse.	Avaliação da leitura realizada pelo OCR.	Demonstrou técnicas complementares de processamento e localização de regiões de interesse.
Uezima et al. (2018)	Utilizando o Tesseract OCR para reconhecimento de textos à mão.	Preparação das imagens para treinamento e reconhecimento textual.	Tesseract OCR.	Reconhecimento de palavras manuscritas.	Distância de Levenshtein e índice de confiança do OCR.	Confirmou a aplicabilidade do Tesseract OCR e a influência da qualidade da imagem na acurácia.
Presente Pesquisa (2026)	Extração automática de informações presentes em RG, CIN e CNH.	Conversão para tons de cinza, binarização, redução de ruídos, segmentação de regiões textuais e múltiplos modos PSM.	Tesseract OCR (PSM 4, 6 e 11).	Extração de campos específicos conforme o tipo documental, incluindo nome, CPF, número do documento, datas, sexo, naturalidade, órgão expedidor, categoria, nome do pai, nome da mãe e número de registro.	Expressões regulares, validação de padrões documentais e comparação entre resultados dos PSMs.	Integra PDI, OCR, extração de campos específicos, validação por regras documentais e validação assistida pelo usuário em documentos oficiais brasileiros.

Fonte: Elaborado pelo autor (2026).

4 METODOLOGIA

Este trabalho adota como metodologia o PDSII, por se tratar de uma abordagem que permite o desenvolvimento gradual da solução por meio de ciclos sucessivos de implementação, testes e refinamentos.

Para apoiar a condução das atividades e a melhoria contínua da solução proposta, foi utilizado o ciclo PDCA como modelo de organização e acompanhamento das etapas do projeto.

Dessa forma, o ciclo PDCA foi utilizado como estrutura organizacional das principais etapas da pesquisa, enquanto o PDSII orientou o desenvolvimento incremental dos módulos da solução.

A combinação dessas abordagens permitiu o desenvolvimento progressivo do sistema, possibilitando a realização de ajustes contínuos a partir dos resultados obtidos durante os testes realizados em ambiente laboratorial controlado.

4.1 Caracterização da pesquisa

Esta pesquisa caracteriza-se como aplicada, exploratória e de caráter experimental. É aplicada por buscar a utilização prática do conhecimento na solução de um problema concreto; exploratória por proporcionar maior familiaridade e esclarecimento sobre o problema investigado; e de caráter experimental por envolver o desenvolvimento e a avaliação de um protótipo mediante testes realizados em ambiente laboratorial controlado (Gil, 1989, p. 43-44, 73-74).

4.2 Construção e aprimoramento com o PDSII

Nesta pesquisa, adotou-se o PDSII, que organiza o desenvolvimento em ciclos sucessivos. Cada ciclo contém um conjunto de funcionalidades desenvolvidas e testadas antes de sua incorporação às etapas seguintes, permitindo ajustes com base nos resultados dos testes laboratoriais e nas inconsistências identificadas durante o desenvolvimento.

Segundo Rezende et al. (2021), o método PDSII compreende, em síntese, quatro etapas: concepção, elaboração, construção e transição. Ademais, os autores

explicam que o PDSII foi utilizado como modelo de desenvolvimento do software e o ciclo PDCA foi utilizado como ferramenta de gestão e organização das etapas da pesquisa. Os autores acrescentam que:

> A dinâmica apresentada permite realimentar a pesquisa, de forma contínua e cíclica, gerando novos requisitos, que por sua vez retornarão às oficinas temáticas, para (re)avaliações. Desse modo, objetiva-se o aprimoramento da plataforma, principalmente por basear-se nas necessidades do público-alvo [...] (Rezende et al., 2021, p. 594).

A utilização do PDSII mostrou-se adequada para esta pesquisa devido à necessidade de realização de múltiplos testes envolvendo processamento de imagens digitais, reconhecimento textual automatizado e validação das informações extraídas.

Durante o desenvolvimento da solução, foi necessário testar diferentes técnicas de pré-processamento de imagens, configurações do OCR e estratégias de validação textual para identificar quais abordagens apresentaram melhor desempenho no reconhecimento das informações presentes nos documentos analisados.

A abordagem iterativa permitiu desenvolver progressivamente o modelo computacional, realizar ajustes nos algoritmos utilizados e aperfeiçoar continuamente os resultados obtidos durante os testes laboratoriais.

No contexto desta pesquisa, a abordagem PDSII foi aplicada por meio do desenvolvimento incremental dos módulos que compõem a solução proposta. Inicialmente foram implementadas as funcionalidades relacionadas ao carregamento das imagens e ao pré-processamento dos documentos.

Em seguida, foram incorporados os mecanismos de reconhecimento óptico de caracteres, extração automatizada dos campos documentais, validação dos dados extraídos e armazenamento dos resultados.

Cada incremento foi submetido a testes individuais antes de sua integração ao sistema completo, permitindo identificar falhas precocemente e promover ajustes contínuos durante o desenvolvimento.

4.3 Utilização do ciclo PDCA

O ciclo PDCA é uma abordagem voltada à organização e à melhoria contínua, composta pelas etapas Plan (Planejar), Do (Executar), Check (Verificar) e Act (Agir). Segundo Rezende et al. (2021), sua utilização em conjunto com o PDSII favorece a

flexibilidade, o controle e o aprimoramento contínuo dos processos e produtos desenvolvidos.

No contexto desta pesquisa, o PDCA foi utilizado como mecanismo organizacional para estruturar o desenvolvimento e a validação do modelo computacional proposto. A metodologia permitiu conduzir o projeto de forma estruturada, permitindo a implementação, avaliação e refinamento contínuo da solução desenvolvida.

A aplicação do PDCA ocorreu em conjunto com o PDSII, especialmente durante as etapas relacionadas ao desenvolvimento do modelo computacional, integração dos módulos e realização dos testes experimentais.

4.4 Desenvolvimento da solução

Com base no PDSII e na utilização do ciclo PDCA como mecanismo de organização da pesquisa, foi estruturado o desenvolvimento da solução proposta. As atividades realizadas durante a construção, avaliação e refinamento do modelo computacional foram distribuídas conforme as etapas do ciclo PDCA, apresentadas no Quadro 2.

Quadro 2 - Distribuição das atividades do ciclo PDCA

Etapa	Atividades
Plan (Planejar)	Levantamento do problema, identificação dos documentos utilizados, definição dos requisitos, seleção das tecnologias e planejamento dos testes.
Do (Executar)	Desenvolvimento do modelo computacional, implementação dos módulos de processamento digital de imagens, OCR, extração e validação dos dados.
Check (Verificar)	Realização de testes laboratoriais, análise dos resultados e avaliação da qualidade das informações extraídas.
Act (Agir)	Realização de ajustes, refinamentos e melhorias a partir dos resultados obtidos durante os testes.

Fonte: Elaborado pelo autor (2026).

4.4.1 Plan (Planejar)

A etapa de planejamento teve como finalidade compreender o problema

investigado e definir os elementos necessários para o desenvolvimento da solução proposta.

Foi realizado um levantamento inicial junto à SRA do IF Baiano – Campus Catu, com o objetivo de compreender o fluxo de cadastramento acadêmico, identificar os documentos mais utilizados e mapear as principais dificuldades relacionadas à inserção manual de dados. Durante essa atividade foi observado que diversas informações presentes em documentos oficiais precisam ser consultadas e inseridas manualmente no SUAP, tornando o processo repetitivo e suscetível a erros de digitação.

A partir desse levantamento foram identificados os documentos mais frequentemente utilizados nos processos de cadastramento, destacando-se o RG, a CIN e a CNH.

Com base nas informações coletadas foram definidos os requisitos funcionais e não funcionais da solução que estão sistematizados nos Quadros 3 e 4.

Quadro 3 - Requisitos funcionais da solução

Código	Requisito
RF01	Permitir o carregamento de imagens dos documentos nas telas correspondentes ao tipo documental selecionado
RF02	Aplicar técnicas de pré-processamento nas imagens
RF03	Executar reconhecimento óptico de caracteres (OCR)
RF04	Identificar automaticamente os campos documentais
RF05	Validar informações extraídas
RF06	Disponibilizar recurso de validação assistida pelo usuário, permitindo conferência, correção e confirmação dos dados extraídos.
RF07	Exportar os dados em formato de valores separados por vírgula (CSV)

Fonte: Elaborado pelo autor (2026).

Quadro 4 - Requisitos não funcionais da solução

Código	Requisitos não funcionais
RNF01	A solução deve ser desenvolvida utilizando a linguagem de programação Python versão 3.10.0
RNF02	A solução deve utilizar a biblioteca OpenCV, versão 4.12.0, para execução das etapas de Processamento Digital de Imagens
RNF03	A solução deve utilizar a biblioteca Tkinter, versão 8.6, para implementação das telas de interação com o usuário.
RNF04	A solução deve funcionar em computadores pessoais com sistema operacional Windows.
RNF05	A solução deve aceitar documentos digitais provenientes de diferentes meios, incluindo imagens nos formatos JPEG e PNG, recebidos por e-mail, produzidos por digitalização ou capturados por dispositivos móveis.
RNF06	A solução deve operar localmente, sem dependência obrigatória de conexão com a internet.
RNF07	A solução deve utilizar o mecanismo Tesseract OCR, versão Tesseract OCR v5.5.0.20241111, para reconhecimento textual.
RNF08	A solução deve permitir exportação dos dados extraídos para arquivo CSV
RNF09	A solução deve processar documentos RG, CIN e CNH.
RNF10	A solução deve processar os documentos localmente, sem envio automático das imagens ou dos dados extraídos para servidores externos ou serviços em nuvem.
RNF11	A solução deve evitar o registro de dados pessoais nos logs, priorizando o armazenamento de eventos técnicos, erros, etapas executadas e status do processamento.
RNF12	A solução deve permitir que os arquivos gerados durante os testes, como CSVs e logs, sejam armazenados apenas pelo tempo necessário à avaliação da pesquisa.

Fonte: Elaborado pelo autor (2026).

Foram definidos os documentos a serem utilizados, conforme o Quadro 5.

Quadro 5 - Campos avaliados por tipo documental

Documento	Campos avaliados
RG	nome, número do RG, data de nascimento, naturalidade, órgão expedidor, CPF e data de emissão.
CIN	nome, sexo, data de nascimento, data de validade, naturalidade, órgão expedidor, CPF e data de emissão.
CNH	nome, data de nascimento, CPF, número da identidade, categoria, nome do pai, nome da mãe, data de emissão, data de validade e número de registro.

Fonte: Elaborado pelo autor (2026).

Em seguida, definiu-se que a avaliação técnica da solução seria realizada em ambiente laboratorial controlado utilizando imagens reais e documentos sintéticos gerados por Inteligência Artificial, permitindo avaliar o desempenho do sistema e reduzir a exposição de dados pessoais de terceiros.

O planejamento dos testes foi estruturado com o objetivo de avaliar, de forma gradual, o comportamento da solução durante todas as etapas do fluxo de processamento. Inicialmente, definiu-se que os testes seriam realizados em ambiente laboratorial controlado, utilizando documentos dos tipos RG, CIN e CNH, por serem os documentos diretamente relacionados ao problema de pesquisa.

Em seguida, foram definidos os campos a serem extraídos conforme o tipo documental. Para o RG, foram considerados nome, número do RG, data de nascimento, naturalidade, órgão expedidor, CPF e data de emissão. Para a CIN, foram considerados nome, sexo, data de nascimento, data de validade, naturalidade, órgão expedidor, CPF e data de emissão. Para a CNH, foram considerados nome, data de nascimento, CPF, número da identidade, categoria, nome do pai, nome da mãe, data de emissão, data de validade e número de registro, conforme sistematizado no Quadro 5.

Também foi planejada a utilização de documentos digitais obtidos por diferentes meios, incluindo fotografias capturadas por dispositivos móveis, imagens nos formatos JPEG e PNG e documentos sintéticos gerados por inteligência artificial. Essa escolha teve como finalidade simular situações próximas ao contexto real de uso da SRA, considerando que os documentos podem ser apresentados com diferentes níveis de qualidade, enquadramento, resolução, iluminação e contraste.

Em seguida, foram definidos testes específicos para cada módulo da solução. Primeiramente, planejou-se testar o carregamento dos arquivos nas telas de validação, verificando se o sistema seria capaz de aceitar os formatos previstos nos requisitos não funcionais. Depois, foram planejados testes das etapas de pré-processamento, com a aplicação de conversão para tons de cinza, binarização e redução de ruídos, a fim de verificar se essas técnicas melhorariam a legibilidade dos caracteres antes da execução do OCR.

Na sequência, foram planejados os testes do mecanismo Tesseract OCR, com a utilização de diferentes modos de segmentação de página (PSM), especialmente os modos 4, 6 e 11, para comparar o comportamento do reconhecimento textual em documentos com diferentes organizações visuais.

Após essa etapa, previu-se a realização de testes das rotinas de extração automática dos campos documentais, utilizando expressões regulares e regras específicas para localizar informações como: CPF; datas; número de RG; órgão expedidor; nome; e demais dados de interesse.

Também foi previsto o teste dos mecanismos de validação dos dados extraídos. Para isso, definiu-se que cada resultado obtido pelo sistema seria comparado manualmente com as informações presentes no documento original.

Os resultados foram classificados como completos, quando a informação fosse extraída corretamente e sem divergências; parciais, quando houvesse erros ou omissões; e falhos, quando a informação não fosse reconhecida ou estivesse incorreta. Essa classificação serviria de base para o cálculo da Taxa de Acerto (TA) e da Taxa de Reconhecimento Aproveitável (TRA), indicadores utilizados para avaliar quantitativamente o desempenho da solução.

Depois, foram planejados testes das telas de validação, da possibilidade de conferência e correção manual dos dados extraídos, da exportação das informações em arquivo CSV e da geração de logs do sistema.

Por fim, o planejamento dos testes não se limitou à verificação do OCR isoladamente, mas contemplou todo o fluxo da solução proposta, desde o carregamento da imagem até a validação, correção, armazenamento e rastreabilidade técnica do processamento.

4.4.2 Do (Executar)

Após a definição dos requisitos funcionais, dos requisitos não funcionais e das tecnologias selecionadas, iniciou-se a implementação do código-fonte da solução proposta. Nessa etapa foram desenvolvidos os algoritmos responsáveis pelo processamento digital das imagens, reconhecimento óptico de caracteres, extração e validação das informações documentais, bem como os componentes das telas de validação e da exportação dos dados.

Inicialmente foram implementadas as funcionalidades responsáveis pelo carregamento das imagens e pelas etapas de pré-processamento. Nessa fase foram utilizadas técnicas de conversão para tons de cinza, binarização e redução de ruídos, com o objetivo de melhorar a qualidade visual das imagens submetidas ao OCR.

Posteriormente foi integrado o mecanismo Tesseract OCR, responsável pela

identificação dos caracteres presentes nos documentos analisados. Em seguida foram desenvolvidas rotinas de extração e validação dos dados reconhecidos, utilizando expressões regulares e regras específicas para identificação de campos do tipo: nome; CPF; número do documento; datas; e demais informações específicas de cada modelo documental, incluindo órgão expedidor quando aplicável.

Também foram desenvolvidas telas de interação com o usuário, implementadas com o auxílio da biblioteca Tkinter, permitindo ao usuário selecionar o tipo documental, carregar o arquivo correspondente, visualizar os dados extraídos e realizar correções quando necessário.

Ao longo do desenvolvimento foram realizados testes de forma gradual para verificar o funcionamento dos módulos implementados e identificar possíveis falhas durante a integração dos componentes.

4.4.3 Check (Verificar)

A etapa de verificação teve como objetivo avaliar tecnicamente o funcionamento da solução desenvolvida.

Os documentos utilizados nos experimentos foram obtidos por diferentes meios, incluindo fotografias capturadas por dispositivos móveis, imagens nos formatos JPEG e PNG e documentos sintéticos gerados com o auxílio da ferramenta Google Gemini. A utilização de documentos sintéticos teve como finalidade exclusiva ampliar a base de testes da pesquisa e evitar a exposição de dados pessoais de terceiros, especialmente por se tratar de documentos oficiais de identificação.

Os documentos sintéticos foram elaborados com dados fictícios, sem correspondência intencional com pessoas reais, e utilizados apenas em ambiente laboratorial controlado. A ferramenta de Inteligência Artificial foi empregada somente como apoio à geração visual desses documentos, permanecendo sob responsabilidade do autor a definição dos procedimentos metodológicos, a seleção das amostras, a execução dos testes, a classificação dos resultados, a análise dos dados e as conclusões apresentadas neste trabalho.

Ressalta-se que os documentos sintéticos não substituem testes em ambiente real de uso, mas permitiram avaliar o comportamento inicial da solução sem comprometer a privacidade de terceiros. Por essa razão, os resultados obtidos devem ser interpretados como uma validação laboratorial do modelo computacional proposto.

Para cada documento processado foi realizada uma comparação manual entre os dados presentes no documento original e as informações extraídas automaticamente pelo sistema.

A avaliação foi realizada campo a campo, tomando-se como referência o conteúdo visualmente presente no documento original e confrontando-o com o resultado inserido automaticamente no campo correspondente pela solução. Para preservar a avaliação do reconhecimento automático, a classificação foi atribuída antes de qualquer correção manual realizada na tela de validação. Em contrapartida, omissões, substituições, acréscimos ou inversões de caracteres que alterassem o conteúdo do dado foram classificados conforme os critérios estabelecidos no Quadro 6.

Para a classificação dos resultados obtidos pelo sistema, foram adotados os critérios apresentados no Quadro 6, considerando a qualidade das informações capturadas automaticamente.

Quadro 6 - Critérios de classificação dos resultados

Classificação	Descrição
Completo	Campo identificado e preenchido automaticamente com conteúdo integralmente correspondente ao valor de referência, sem necessidade de correção manual.
Parcial	Campo identificado e preenchido automaticamente, mas com omissão, substituição, acréscimo ou inversão de um ou mais caracteres, conservando conteúdo suficiente para permitir a identificação do dado e sua posterior correção pelo usuário.
Falho	Campo não identificado, deixado vazio, associado a informação incorreta ou preenchido com conteúdo que não permita identificar com segurança o dado presente no documento original.

Fonte: Elaborado pelo autor (2026).

A classificação em completo, parcial e falho foi adotada porque a unidade de análise desta pesquisa corresponde a cada campo documental estruturado, como nome, CPF, data de nascimento e número do documento. Essa classificação permite distinguir três situações relevantes para o funcionamento do protótipo: a extração integral do dado; a extração incompleta, mas ainda identificável e passível de correção; e a ausência de conteúdo utilizável. A escolha também se relaciona ao modelo de validação assistida empregado na solução, no qual o resultado automático é apresentado ao usuário para conferência antes de seu armazenamento.

Para fins de avaliação quantitativa da solução, foram definidos, no âmbito desta

pesquisa, dois indicadores:

a) Taxa de Acerto (TA): TA = (Quantidade de campos classificados como completos ÷ Quantidade total de campos avaliados) × 100
b) Taxa de Reconhecimento Aproveitável (TRA): TRA = ((Quantidade de campos completos + Quantidade de campos parciais) ÷ Quantidade total de campos avaliados) × 100

A Taxa de Acerto (TA) foi escolhida como indicador principal por representar um critério estrito de reconhecimento: somente são contabilizados como acertos os campos extraídos integralmente, sem divergência em relação ao conteúdo de referência e sem necessidade de correção manual. Dessa forma, a TA permite mensurar a proporção de campos que poderiam ser confirmados pelo usuário sem alteração de seu conteúdo.

A Taxa de Reconhecimento Aproveitável (TRA) foi adotada como indicador complementar em razão da natureza assistiva da solução desenvolvida. Como o protótipo não foi concebido para substituir a conferência humana, determinados campos parcialmente reconhecidos, embora não constituam acertos integrais, podem conservar conteúdo suficiente para auxiliar a identificação e a correção do dado pelo usuário. A TRA, portanto, reúne os resultados completos e parciais para demonstrar a proporção de campos nos quais houve algum conteúdo reconhecido com possibilidade de aproveitamento durante a validação assistida.

A TRA não deve ser interpretada como taxa de exatidão absoluta, nem como demonstração de que o campo poderia ser utilizado sem revisão. Também não comprova redução de tempo, produtividade, diminuição de erros de digitação ou economia de esforço operacional, pois essas dimensões não foram mensuradas nesta pesquisa. Os dois indicadores foram escolhidos para representar, respectivamente, o reconhecimento integral dos campos e o potencial de aproveitamento dos resultados no fluxo de conferência humana adotado pelo protótipo.

Os resultados foram analisados por tipo documental, considerando a distribuição dos campos entre as três classificações e os valores obtidos para a TA e a TRA. Essa abordagem possibilitou identificar diferenças no desempenho da solução entre RG, CIN e CNH, bem como limitações relacionadas à qualidade das imagens, à diversidade de layouts, às regras de extração e às configurações empregadas no reconhecimento textual.

4.4.4 Act (Agir)

Com base nos resultados obtidos que se encontram na Seção 6, durante os testes foram realizados ajustes nas técnicas de pré-processamento e nos parâmetros do mecanismo OCR.

Também foram refinadas as regras utilizadas para identificação e validação dos campos documentais, especialmente nos campos relacionados a CPF, datas e números de documentos.

Além disso, foram efetuadas melhorias nas telas de validação para facilitar a conferência das informações extraídas e corrigir inconsistências identificadas durante os testes.

Os refinamentos realizados foram incorporados à versão final do protótipo, abrangendo ajustes no pré-processamento, no reconhecimento textual, nas regras de extração e nas telas de validação. Como não foi realizada comparação quantitativa entre as versões anteriores e a versão final, não se pretende atribuir a esses ajustes ganho mensurado de precisão ou estabilidade.

As mesmas amostras foram utilizadas nos ciclos de desenvolvimento, ajuste e avaliação final do protótipo. Dessa forma, os resultados apresentados possuem caráter exploratório e podem refletir adaptação da implementação às características da base experimental utilizada. Não foi empregado um conjunto independente de teste, circunstância que constitui uma limitação metodológica da pesquisa.

4.5 Cuidados éticos e proteção de dados pessoais

Considerando que os documentos analisados nesta pesquisa contêm dados pessoais, como: nome; CPF; data de nascimento; filiação; número de documento; e demais informações de identificação civil, foram adotados cuidados metodológicos voltados à proteção dessas informações durante a realização dos experimentos.

A solução desenvolvida foi executada em ambiente local, sem envio automático das imagens processadas ou dos dados extraídos para serviços externos, servidores remotos ou ambientes em nuvem. O uso de conexão com a internet não foi necessário para a execução do fluxo principal da solução, que compreende o carregamento da imagem, o pré-processamento, a aplicação do OCR, a extração dos campos, a conferência manual e a exportação dos resultados.

Para reduzir a exposição de dados reais, parte da base experimental foi composta por documentos sintéticos gerados com dados fictícios, sem correspondência intencional com pessoas reais. Esses documentos foram utilizados exclusivamente para fins de teste laboratorial da solução, permitindo avaliar o comportamento do modelo computacional sem comprometer a privacidade de terceiros.

As figuras inseridas no trabalho foram selecionadas e tratadas de forma a evitar a divulgação de dados pessoais reais. Quando necessário, os dados pessoais são ocultados, substituídos ou representados por dados fictícios. Além disso, os dados extraídos durante os testes foram utilizados apenas para fins de avaliação acadêmica da solução, não sendo integrados ao SUAP nem utilizados para cadastramento real de estudantes.

Também se buscou limitar o armazenamento das informações processadas. Os arquivos CSV e os registros de log foram utilizados apenas como instrumentos de validação, rastreabilidade técnica do processamento e análise dos resultados da pesquisa. Recomenda-se que, em eventual implantação futura da solução em ambiente institucional, sejam adotadas políticas formais de controle de acesso, retenção mínima de dados, descarte seguro, autenticação de usuários e adequação integral às normas de proteção de dados pessoais aplicáveis.

Concluída a etapa metodológica, iniciou-se o desenvolvimento do modelo computacional proposto. A seção seguinte apresenta a arquitetura da solução, as técnicas de Processamento Digital de Imagens empregadas, os mecanismos de reconhecimento óptico de caracteres utilizados e as estratégias adotadas para extração, validação e armazenamento das informações presentes nos documentos analisados.

5 MODELO COMPUTACIONAL

O modelo computacional apresentado nesta seção foi desenvolvido considerando os requisitos funcionais e não funcionais definidos na etapa de planejamento da pesquisa, especialmente os requisitos relacionados ao carregamento dos documentos, aplicação de técnicas de Processamento Digital de Imagens, reconhecimento óptico de caracteres, validação dos dados extraídos e disponibilização dos resultados ao usuário.

5.1 Arquitetura da solução

Antes da apresentação detalhada dos módulos implementados, torna-se importante compreender a arquitetura geral da solução desenvolvida. O modelo computacional proposto foi estruturado como um pipeline de processamento composto por etapas sequenciais, nas quais a saída de cada módulo serve como entrada para a etapa seguinte.

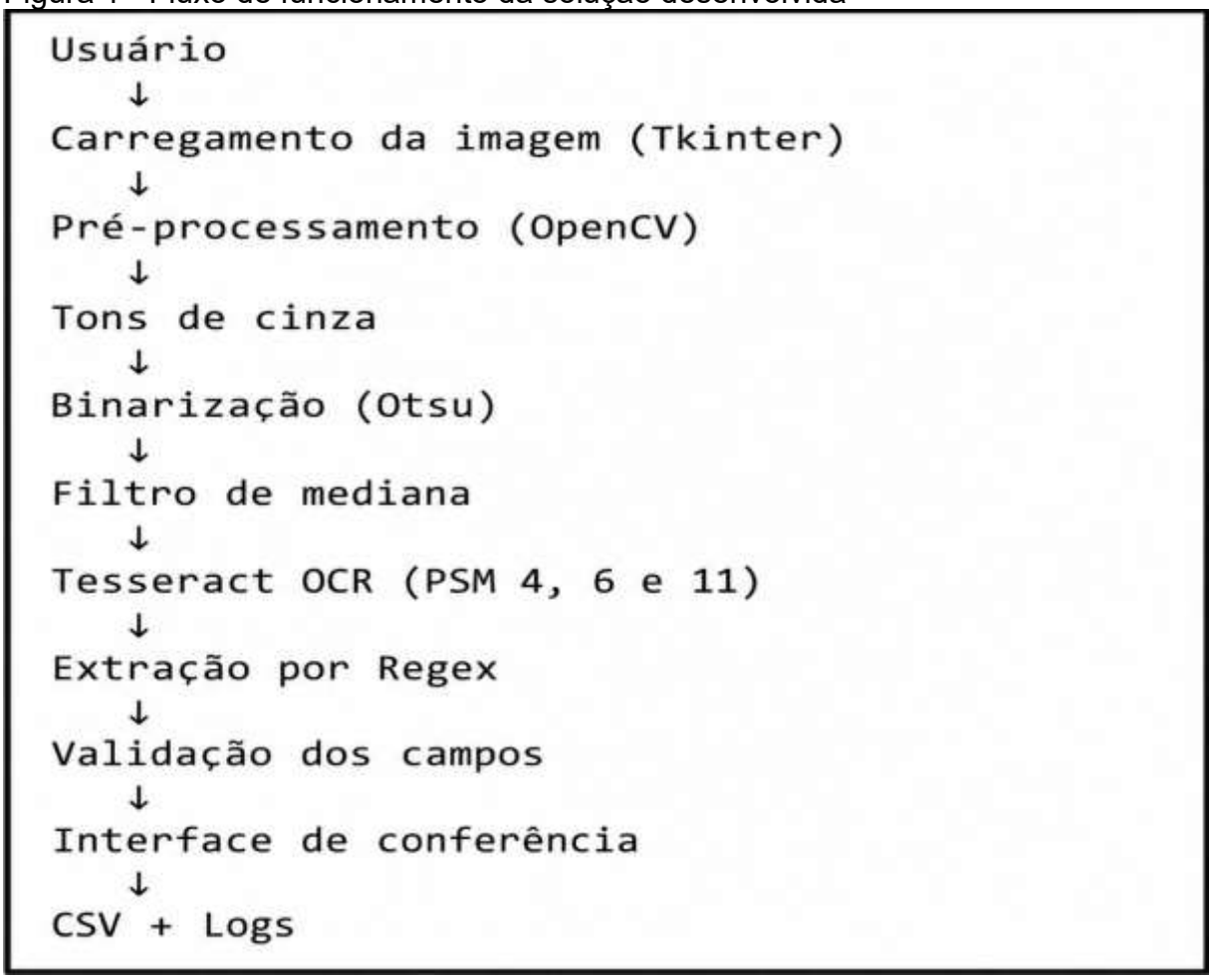
Inicialmente, o usuário seleciona, na tela inicial da aplicação, o tipo documental a ser processado, escolhendo entre CNH ou RG/CIN. Após essa escolha, a aplicação direciona o usuário para a tela de carregamento e validação correspondente ao tipo documental selecionado. Nessa tela, o usuário seleciona o arquivo contendo o documento a ser analisado, dando início ao fluxo de processamento.

Em seguida, a imagem é submetida às técnicas de Processamento Digital de Imagens, incluindo conversão para tons de cinza, binarização pelo método de Otsu e redução de ruídos utilizando o filtro de mediana. Após o pré-processamento, o mecanismo Tesseract OCR realiza o reconhecimento dos caracteres presentes no documento.

O texto obtido pelo OCR é então submetido a rotinas de extração baseadas em expressões regulares e regras documentais específicas para cada modelo de documento. Por fim, os dados extraídos são apresentados na própria tela de carregamento e validação, permitindo conferência manual e correção pelo usuário antes do armazenamento em arquivo CSV e do registro dos eventos de execução e rastreabilidade técnica do processamento.

A Figura 1 apresenta uma visão geral do fluxo de funcionamento da solução desenvolvida.

Figura 1 - Fluxo de funcionamento da solução desenvolvida



Fonte: Elaborado pelo autor (2026).

5.2 Aquisição e imagens

A etapa de aquisição das imagens representa o ponto inicial do fluxo de processamento da solução computacional desenvolvida neste trabalho. A qualidade das imagens utilizadas como entrada possui influência direta sobre o desempenho das etapas subsequentes de PDI, OCR e extração automatizada das informações.

Considerando que os documentos apresentados pelos estudantes durante o processo de matrícula podem ser obtidos por diferentes meios, a solução foi projetada para aceitar imagens oriundas de fotografias realizadas por dispositivos móveis e arquivos digitais previamente armazenados em formatos compatíveis.

Com o objetivo de simular cenários próximos à realidade da SRA, foram utilizadas imagens de documentos de identificação contendo diferentes características de captura, incluindo variações de resolução, iluminação, contraste, enquadramento e posicionamento do documento. Essa diversidade permitiu avaliar a capacidade do

sistema em lidar com situações frequentemente encontradas em ambientes reais de atendimento.

Para preservar a privacidade dos dados pessoais e atender aos princípios éticos relacionados ao tratamento dessas informações, parte das imagens utilizadas nos experimentos foi produzida por meio da geração de documentos sintéticos com auxílio de ferramentas de Inteligência Artificial. Essa abordagem permitiu reproduzir a estrutura visual de documentos oficiais, como RG, CIN e CNH, sem a utilização de dados reais de terceiros.

As Figuras 2, 3 e 4 são exemplos dos modelos de documentos empregados durante os testes do sistema. A Figura 2 demonstra um modelo de referência da CNH, utilizado para representar a organização visual dos campos presentes no documento. As Figuras 3 e 4[2] são documentos sintéticos gerados artificialmente, utilizados para ampliar o conjunto de testes e avaliar o desempenho do algoritmo em diferentes configurações visuais.

Figura 2 - Modelo de CNH

Fonte: UOL Carros, imagem de divulgação, conforme Thomas (2023).

[2] As imagens apresentadas possuem caráter exclusivamente ilustrativo, sendo utilizadas como referência para demonstrar a disposição dos campos textuais que podem ser submetidos ao processo de extração automática de dados.

A utilização de diferentes tipos documentais e diferentes condições de aquisição das imagens foi fundamental para avaliar a capacidade de adaptação da solução frente a variações de resolução, iluminação, enquadramento e qualidade visual dos documentos utilizados nos testes.

Figura 3 - Exemplo de Carteira de Identidade Nacional

Fonte: Documento sintético gerado pelo autor com o auxílio da ferramenta Google Gemini (2025).

Figura 4 - Exemplo de documento sintético de CNH

Fonte: Documento sintético gerado pelo autor com o auxílio da ferramenta Google Gemini (2025).

5.3 Seleção do tipo documental e carregamento das imagens

Após a etapa de aquisição, o usuário inicia o fluxo da aplicação por meio da tela inicial, na qual seleciona o tipo documental a ser processado. A aplicação disponibiliza duas opções principais: uma destinada aos documentos do tipo CNH e outra destinada aos documentos do tipo RG ou CIN.

Essa separação foi adotada porque os documentos analisados possuem estruturas e campos distintos. A CNH apresenta informações específicas relacionadas à habilitação, como número de registro, categoria e data de validade, enquanto RG e CIN concentram campos de identificação civil, como nome, CPF, data de nascimento, filiação, naturalidade e demais informações conforme o modelo documental.

Após a escolha do tipo documental, o usuário é direcionado para a tela de carregamento e validação correspondente ao documento selecionado. Nessa tela, ocorre a seleção do arquivo que será processado, podendo ser utilizado documento em formato de imagem ou arquivo digital compatível com os requisitos definidos para a solução.

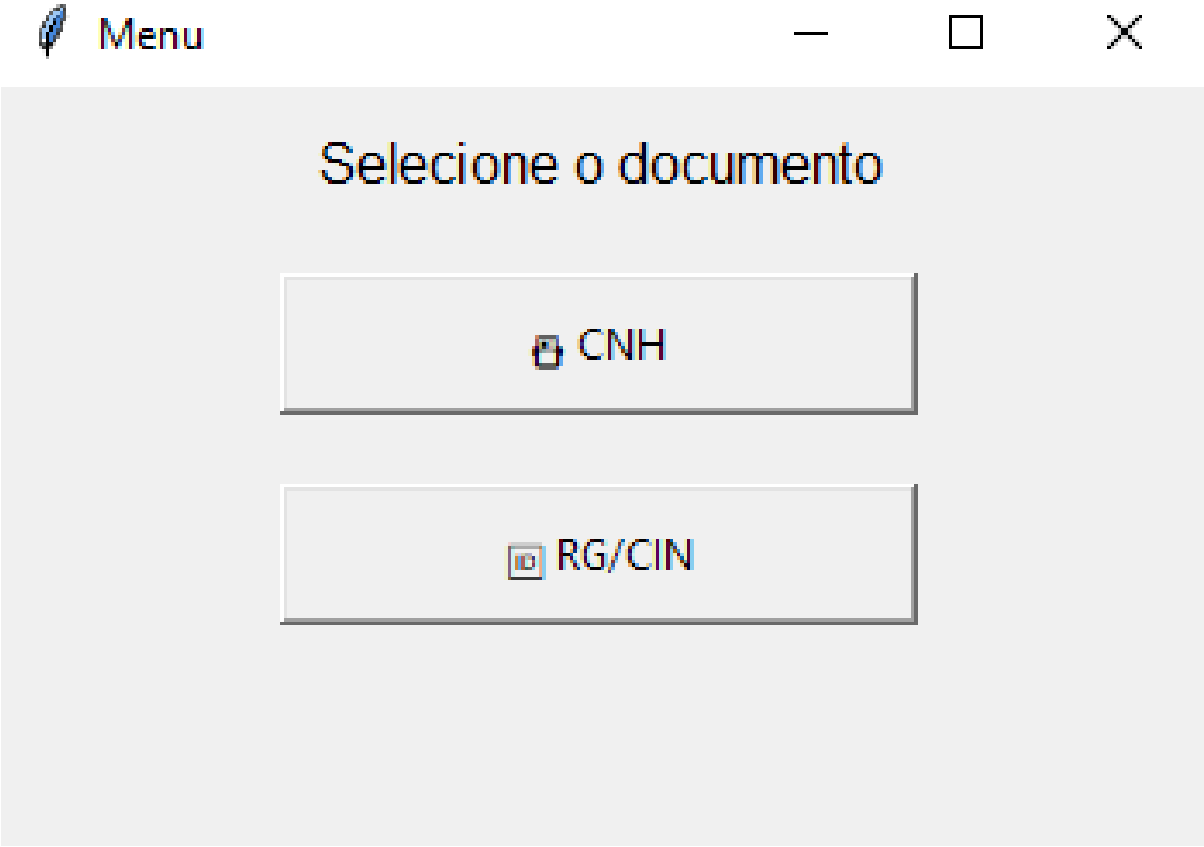
A tela inicial da aplicação e as telas de carregamento e validação foram implementadas com o auxílio da biblioteca Tkinter, conforme definido nos requisitos não funcionais da solução. Neste trabalho, o Tkinter é tratado como recurso de implementação das telas de interação com o usuário, e não como etapa do processamento das imagens.

O carregamento das imagens ocorre a partir da seleção realizada pelo usuário na tela correspondente ao tipo documental escolhido. Após essa seleção, a aplicação realiza a leitura da imagem utilizando os recursos da biblioteca OpenCV, convertendo o arquivo selecionado em uma matriz de pixels que poderá ser manipulada pelas etapas posteriores do processamento.

O processo de carregamento também contempla mecanismos de verificação da existência e da compatibilidade do arquivo selecionado, evitando que imagens inválidas, inexistentes ou incompatíveis sejam encaminhadas às etapas seguintes do fluxo. Essa validação inicial contribui para a estabilidade da aplicação e reduz a ocorrência de falhas durante a execução do algoritmo.

A Figura 5 apresenta a tela inicial de seleção do tipo documental, utilizada para direcionar o usuário à tela correspondente ao documento que será processado.

Figura 5 - Tela inicial de seleção do tipo documental



Fonte: Elaborado pelo autor (2026).

As telas de validação foram desenvolvidas apresentando ao usuário apenas os elementos necessários para a execução do processo, como a seleção da imagem, a visualização do documento carregado, a exibição dos dados extraídos e as opções relacionadas ao armazenamento dos resultados.

5.4 Pré-processamento das imagens

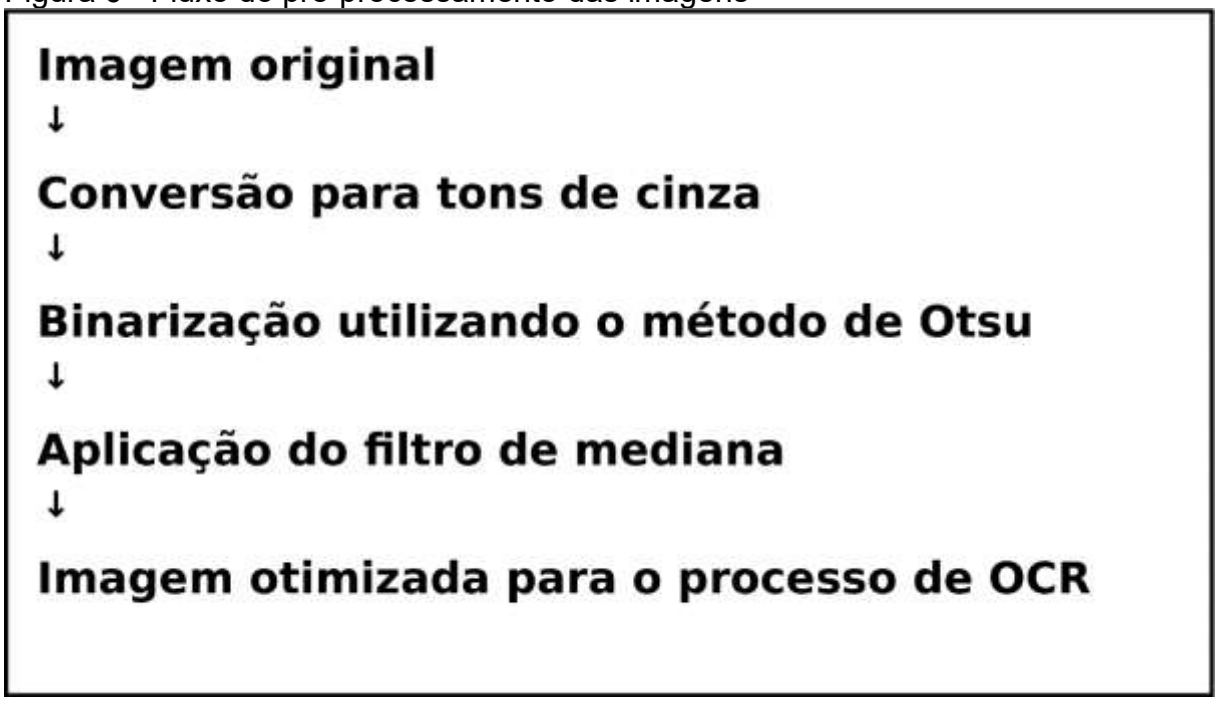
Após o carregamento do documento, inicia-se a etapa de pré-processamento das imagens, responsável por preparar os dados de entrada para o mecanismo de Reconhecimento Óptico de Caracteres. Essa etapa possui papel essencial no desempenho do sistema, pois a precisão do OCR depende diretamente da qualidade visual das imagens analisadas.

Conforme discutido na Seção 2, o Processamento Digital de Imagens permite aplicar técnicas capazes de reduzir interferências visuais, melhorar o contraste dos caracteres e destacar informações relevantes presentes no documento. Dessa forma, antes da execução do OCR, as imagens passam por uma sequência de

transformações destinadas a aumentar a legibilidade dos elementos textuais.

O fluxo de pré-processamento implementado no modelo computacional segue a sequência apresentada na Figura 6.

Figura 6 - Fluxo do pré-processamento das imagens



Fonte: Elaborado pelo autor (2026).

A escolha dessa sequência de processamento foi baseada nas características dos documentos analisados e nos fundamentos apresentados na revisão bibliográfica.

A conversão para tons de cinza reduz a complexidade da informação visual, a binarização aumenta a separação entre caracteres e plano de fundo, enquanto a filtragem por mediana reduz ruídos que poderiam interferir no reconhecimento textual. A aplicação dessas técnicas em conjunto permite que a imagem enviada ao OCR apresente maior qualidade e uniformidade, buscando preparar as imagens e favorecer o reconhecimento dos caracteres dos campos documentais. Esse procedimento é especialmente importante em documentos fotografados por dispositivos móveis, nos quais são comuns problemas como sombras, reflexos, pequenas inclinações e variações de iluminação.

Portanto, o pré-processamento representa uma etapa estratégica dentro da arquitetura da solução desenvolvida, atuando como um mecanismo de preparação dos dados antes da aplicação dos algoritmos de reconhecimento textual.

5.5 Conversão para tons de cinza

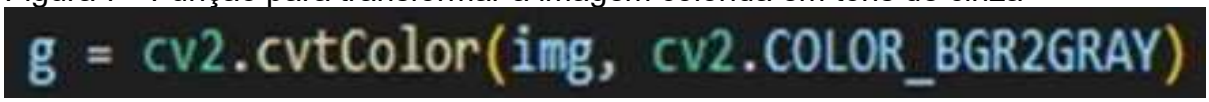
A primeira transformação aplicada durante o pré-processamento das imagens consiste na conversão do documento original em uma imagem em escala de cinza. Embora os documentos de identificação possuam diferentes elementos visuais coloridos, como fotografias, brasões, marcas de segurança e fundos com padrões gráficos, grande parte das informações relevantes para o objetivo desta pesquisa encontra-se representada por caracteres textuais, os quais dependem principalmente da variação de intensidade luminosa entre o texto e o fundo do documento.

Dessa forma, a conversão para tons de cinza foi adotada como uma estratégia para reduzir a quantidade de informações processadas pelo sistema, eliminando dados cromáticos que não contribuem diretamente para o reconhecimento dos caracteres. Essa redução da complexidade da imagem permite diminuir o custo computacional das etapas seguintes, tornando o processamento mais eficiente.

No modelo desenvolvido, essa conversão foi realizada utilizando a função cv2.cvtColor() da biblioteca OpenCV, responsável pela conversão da representação de cores azul, verde e vermelho (BGR), utilizada pela OpenCV, para uma representação monocromática em tons de cinza, baseada nos níveis de intensidade dos pixels.

A Figura 7 apresenta o trecho responsável pela conversão da imagem original para tons de cinza.

Figura 7 - Função para transformar a imagem colorida em tons de cinza



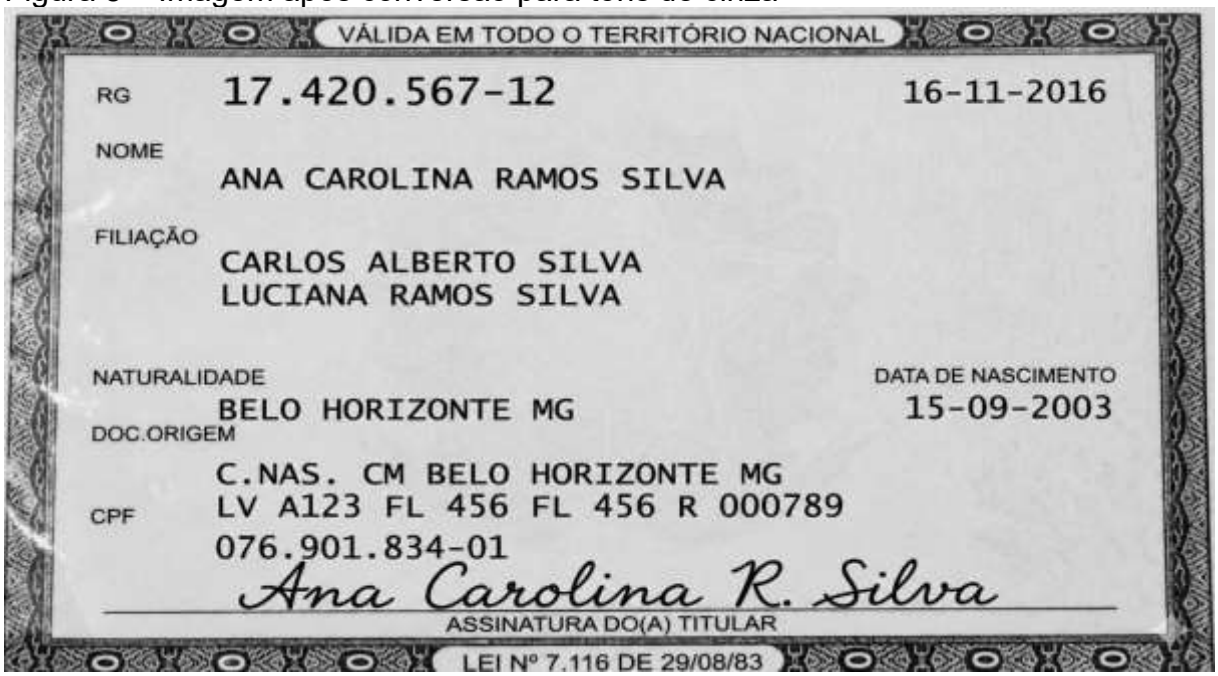
Fonte: Elaborado pelo autor (2026).

A variável img representa a imagem obtida durante a etapa de carregamento, enquanto o parâmetro cv2.COLOR_BGR2GRAY determina que a conversão seja realizada a partir do padrão de cores utilizado internamente pela biblioteca OpenCV para uma representação em escala de cinza.

A escolha dessa técnica está diretamente relacionada às características dos documentos analisados, uma vez que a remoção das informações de cor contribui para destacar os caracteres e preparar a imagem para a etapa posterior de segmentação por binarização.

A utilização dessa etapa, além de reduzir a quantidade de informações processadas, contribui para padronizar imagens obtidas em diferentes condições de captura, com o objetivo de reduzir os efeitos das variações de iluminação, tonalidade e características visuais dos documentos utilizados durante os testes.

Figura 8 – Imagem após conversão para tons de cinza



Fonte: Elaborado pelo autor (2026), a partir da aplicação do código desenvolvido a documento sintético gerado com auxílio do Google Gemini (2025).

5.6 Segmentação da imagem pelo método de Otsu

Após a conversão da imagem para tons de cinza, o próximo estágio do pré-processamento consiste na aplicação da técnica de segmentação por binarização. Essa etapa possui como objetivo separar os caracteres presentes no documento do plano de fundo, aumentando o contraste entre as regiões de interesse e facilitando a atuação do mecanismo de reconhecimento óptico de caracteres.

Entre os diferentes métodos de limiarização disponíveis, optou-se pela utilização do método de Otsu devido à sua capacidade de determinar o valor de limiar calculado automaticamente para cada imagem processada. Essa característica é especialmente importante em um cenário como o desta pesquisa, no qual os documentos podem apresentar diferentes condições de iluminação, contraste e qualidade de captura.

A implementação da binarização utilizando o método de Otsu foi realizada por

meio da função cv2.threshold() da biblioteca OpenCV, conforme apresentado na Figura 9.

Figura 9 – Trecho de código para binarização pelo método de Otsu

```
_, b = cv2.threshold(g, 0, 255, cv2.THRESH_BINARY + cv2.THRESH_OTSU)
```

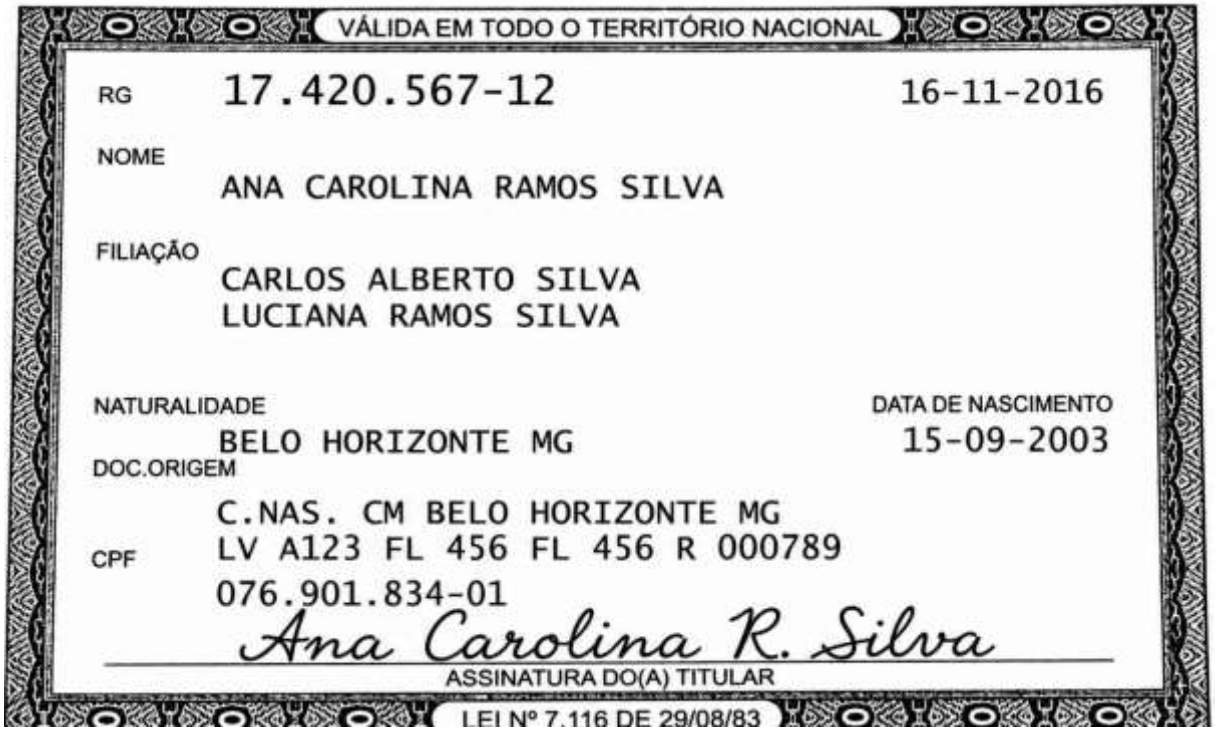
Fonte: Elaborado pelo autor (2026).

Nesse procedimento, o algoritmo analisa a distribuição dos níveis de intensidade dos pixels da imagem em escala de cinza e calcula automaticamente o limiar mais adequado para separar os elementos textuais do fundo.

A variável “b” corresponde ao resultado da segmentação, na qual os pixels são representados apenas por dois valores possíveis: preto e branco. Essa simplificação permite evidenciar os caracteres presentes no documento, reduzindo interferências causadas por fundos coloridos, marcas gráficas e pequenas variações de luminosidade.

A Figura 10 apresenta um exemplo do resultado obtido após a aplicação da binarização pelo método de Otsu.

Figura 10 – Imagem após binarização pelo método de Otsu



Fonte: Elaborado pelo autor (2026), a partir da aplicação do código desenvolvido a documento sintético gerado com auxílio do Google Gemini (2025).

O método de Otsu foi adotado por calcular automaticamente o limiar de binarização de cada imagem, dispensando a definição manual de um valor fixo para todas as amostras. Essa característica mostrou-se compatível com a diversidade de condições visuais presente na base experimental.

5.7 Filtro de mediana (medianBlur)

Após a etapa de binarização, as imagens podem apresentar pequenas imperfeições resultantes do processo de captura ou da própria transformação realizada durante o processamento, tais como pontos isolados, granulações e ruídos visuais. Esses elementos indesejados podem prejudicar o desempenho do OCR, provocando interpretações incorretas dos caracteres.

Para reduzir essas interferências, foi utilizado o filtro de mediana, implementado por meio da função cv2.medianBlur() da biblioteca OpenCV. A escolha desse filtro ocorreu devido à sua capacidade de remover ruídos sem comprometer significativamente as bordas dos caracteres, característica essencial em aplicações de reconhecimento textual.

A Figura 11 apresenta o trecho utilizado para aplicação do filtro de mediana.

Figura 11 - Função utilizada para remoção de ruídos

```
limpa = cv2.medianBlur(b, 3)
```

Fonte: Elaborado pelo autor (2026).

No exemplo apresentado, o primeiro parâmetro corresponde à imagem resultante da etapa de binarização, enquanto o segundo parâmetro define o tamanho da janela utilizada para o cálculo da mediana dos pixels vizinhos. A utilização de uma janela de tamanho reduzido foi escolhida para equilibrar a remoção de ruídos e a preservação dos detalhes dos caracteres.

A utilização do filtro de mediana representa a última etapa do pré-processamento antes do envio da imagem ao mecanismo OCR. Dessa forma, o documento entregue ao Tesseract apresenta maior uniformidade visual e menor quantidade de interferências, aumentando a probabilidade de reconhecimento correto dos campos textuais.

A combinação das etapas de conversão para tons de cinza, binarização pelo método de Otsu e filtragem por mediana constitui um pipeline de preparação das imagens que busca minimizar problemas comuns em documentos capturados em ambientes reais, como variações de iluminação, presença de ruídos e diferenças na qualidade dos dispositivos utilizados para a captura.

Figura 12 – Imagem binarizada após aplicação do filtro de mediana

Fonte: Elaborado pelo autor (2026), a partir da aplicação do código desenvolvido a documento sintético gerado com auxílio do Google Gemini (2025).

Após a conclusão das etapas de pré-processamento, a imagem encontra-se preparada para ser submetida ao mecanismo de Reconhecimento Óptico de Caracteres. A etapa seguinte descreve a integração da solução desenvolvida com o Tesseract OCR, responsável pela conversão das informações visuais presentes no documento em dados textuais que poderão ser posteriormente tratados, validados e armazenados pelo sistema.

5.8 Reconhecimento óptico de caracteres

Após a conclusão das etapas de pré-processamento, a imagem tratada é submetida ao mecanismo de Reconhecimento Óptico de Caracteres (OCR),

responsável por identificar os caracteres presentes no documento e convertê-los em uma representação textual manipulável pela aplicação.

Conforme discutido na Fundamentação Teórica, o OCR constitui uma tecnologia baseada em algoritmos capazes de reconhecer padrões visuais presentes em imagens e convertê-los em caracteres digitais. No modelo desenvolvido neste trabalho, foi utilizado o Tesseract OCR, integrado à linguagem Python.

A escolha do Tesseract ocorreu em razão de suas características técnicas e práticas, destacando-se sua natureza de código aberto, ampla utilização em pesquisas acadêmicas, suporte a múltiplos idiomas e capacidade de integração com bibliotecas de Visão Computacional, como o OpenCV.

O processo de reconhecimento é iniciado pelo envio da imagem previamente tratada ao mecanismo OCR. O Tesseract analisa a estrutura visual da imagem, identifica regiões contendo texto, realiza a segmentação das linhas e palavras e, por meio de modelos treinados de reconhecimento de padrões, converte os caracteres identificados em uma sequência textual.

A Figura 13 apresenta o fluxo simplificado do processo de reconhecimento textual realizado pelo sistema.

Figura 13 - Fluxo do reconhecimento óptico de caracteres

Fonte: Elaborado pelo autor (2026).

O resultado obtido nessa etapa corresponde a um texto bruto, que ainda pode apresentar inconsistências decorrentes das limitações naturais do OCR, como caracteres incorretos, espaços inadequados ou informações duplicadas. Por esse motivo, são necessárias etapas posteriores de interpretação e tratamento dos dados extraídos.

5.9 Configuração dos modos de segmentação de página (PSM)

O desempenho do Tesseract OCR depende não apenas da qualidade da imagem fornecida como entrada, mas também da configuração adequada dos seus parâmetros internos. Um dos principais parâmetros utilizados neste trabalho corresponde ao Page Segmentation Mode (PSM), responsável por informar ao mecanismo OCR como a estrutura da imagem deve ser interpretada.

Como os documentos analisados possuem diferentes formatos e distribuições de campos textuais, não existe uma única configuração de segmentação capaz de apresentar o melhor desempenho em todos os cenários. Dessa forma, a solução foi desenvolvida utilizando diferentes modos de segmentação do Tesseract, permitindo maior flexibilidade na identificação das informações.

Os modos utilizados foram os PSM 4, PSM 6 e PSM 11.

a) PSM 4 – coluna única de texto com tamanhos variáveis: esse modo pressupõe que a imagem contém uma única coluna textual, cujos elementos podem apresentar tamanhos diferentes. Essa configuração mostrou-se útil em documentos que apresentam campos organizados verticalmente, como listas de informações pessoais distribuídas em áreas específicas;
b) PSM 6 – Bloco uniforme de texto: o modo PSM 6 assume que a imagem contém um único bloco uniforme de texto. Esse modo apresentou bom desempenho em regiões de documentos nas quais os campos encontram-se próximos e seguem uma organização linear;
c) PSM 11 – Texto esparso: o modo PSM 11 é indicado para imagens que contêm elementos textuais distribuídos de maneira não uniforme, permitindo que o Tesseract procure textos em diferentes regiões da imagem sem assumir uma estrutura rígida.

A utilização combinada destes modos foi uma decisão estratégica do projeto, pois permitiu comparar diferentes interpretações da mesma imagem, aumentando a

possibilidade de recuperar informações que poderiam não ser reconhecidas adequadamente por apenas uma configuração.

A Figura 14 apresenta um exemplo de configuração do PSM no Tesseract.

Figura 14 - Configuração do idioma e do modo de segmentação no Tesseract

```
# Configs do tesseract
cfgs = ["--psm 6 --oem 3", "--psm 4 --oem 3", "--psm 8 --oem 3"]
outs = []
self._log("=== OCR TESSERACT ===")
for i, c in enumerate(cfgs):
    try:
        self._log(f"Configuração {i+1}: {c}")
        t = pytesseract.image_to_string(limpa, lang="por", config=c)
        outs.append(t)
        self._log(f"Texto extraído (config {i+1}):\n{t}\n{'-' * 40}")
    except Exception as e:
        self._log(f"ERRO na configuração {i+1}: {e}")

txt = max(outs, key=len) if outs else ""
self._log("=== TEXTO SELECIONADO (MAIS COMPLETO) ===")
self._log(txt)
self._log("=" * 50)
```

Fonte: Elaborado pelo autor (2026).

No exemplo apresentado, o parâmetro --oem 3 determina que o Tesseract utilize o modo de mecanismo padrão, selecionado conforme os componentes disponíveis. O parâmetro --psm 6, por sua vez, orienta o mecanismo a tratar a imagem como um único bloco uniforme de texto.

A utilização de múltiplas estratégias de segmentação está diretamente relacionada ao objetivo deste trabalho de aumentar a robustez da solução frente às diferentes características dos documentos analisados.

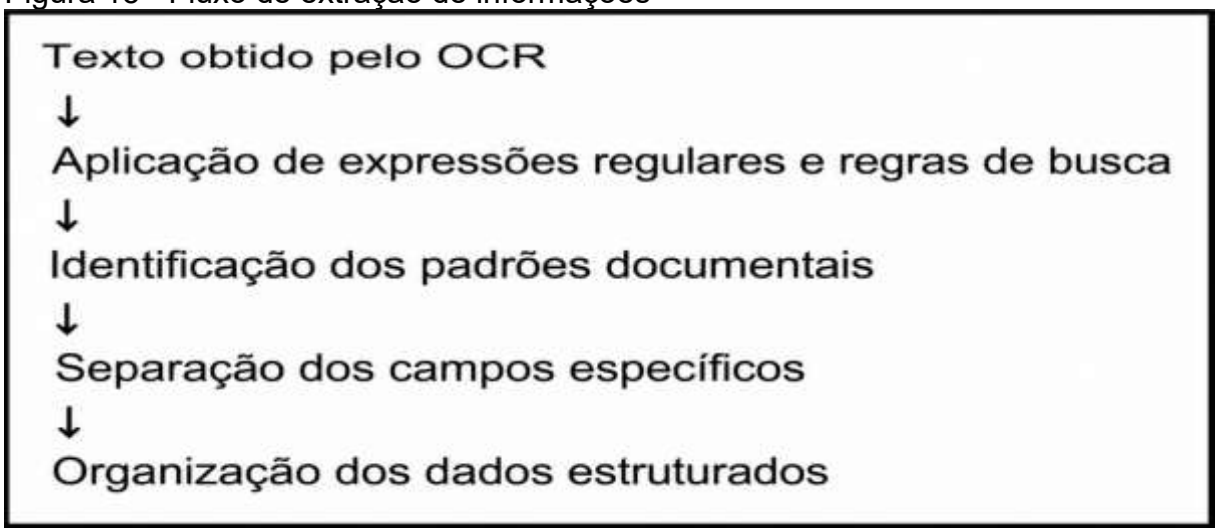
5.10 Extração de dados

O texto retornado pelo OCR corresponde a uma representação digital dos caracteres encontrados na imagem. Entretanto, o objetivo da solução desenvolvida não consiste apenas em transformar imagens em texto, mas em identificar e organizar automaticamente os campos relevantes para o processo de matrícula acadêmica.

Assim, após a obtenção do texto bruto, é executada uma etapa de extração estruturada dos dados por meio de técnicas baseadas em padrões textuais e regras documentais.

O fluxo de extração de informações ocorre conforme apresentado a seguir:

Figura 15 - Fluxo de extração de informações



Fonte: Elaborado pelo autor (2026).

A extração dos campos foi realizada por meio de expressões regulares (Regular Expressions — Regex) e de regras específicas para cada tipo documental, permitindo localizar padrões correspondentes ao CPF, ao número do documento, às datas e às demais informações previstas no Quadro 5. As expressões regulares possibilitam identificar sequências de caracteres que seguem padrões previamente definidos, mesmo quando apresentam variações de formatação no texto reconhecido pelo OCR.

Os campos extraídos pelo sistema foram definidos na etapa Plan, conforme o Quadro 5, apresentado no item 4.4.1.

Por exemplo, o CPF possui uma estrutura padronizada composta por onze dígitos, podendo ser representado com ou sem caracteres de formatação. Dessa forma, foram desenvolvidas expressões regulares capazes de localizar essas variações no texto reconhecido pelo OCR.

Além da identificação por expressões regulares, foram utilizadas regras de associação contextual para relacionar os valores encontrados aos respectivos campos, distinguindo, por exemplo, as diferentes datas presentes no documento, os nomes do pai e da mãe, a categoria da habilitação e o número de registro.

A Figura 16 apresenta um exemplo da função responsável por coordenar as operações de extração dos campos documentais.

Figura 16 - Função que coordena as operações de extração

```
def pegar_dados(self, txt):
    Orquestra todas as funções de extração.
    """
    self._log("=== EXTRAINDO TODOS OS DADOS ===")
    linhas = txt.split("\n")
    self._log(f"Total de linhas detectadas: {len(linhas)}")
    for i, linha in enumerate(linhas):
        self._log(f"Linha {i}: {linha}")

    dados = {}

    # Nome
    nome = self.procurar_nome(linhas, txt)
    if nome:
        dados["nome"] = nome

    # Datas
    datas = re.findall(r"\d{2}/\d{2}/\d{4}", txt)
    if datas:
        dados.update(self.organizar_datas_real(datas, linhas))

    # CPF
    cpf = self.procurar_cpf(txt)
    if cpf:
        dados["cpf"] = cpf

    # Identidade com validação melhorada
    ident = self.procurar_identidade_melhorada(txt, linhas)
    if ident:
        dados["identidade"] = ident
```

Fonte: Elaborado pelo autor (2026).

A utilização de expressões regulares e de regras específicas para cada tipo documental permitiu transformar o texto não estruturado fornecido pelo OCR em campos organizados, compatíveis com as necessidades do processo de cadastramento acadêmico.

5.11 Pós-processamento dos campos extraídos

Embora o OCR seja capaz de converter imagens em texto, os campos posteriormente extraídos podem conter inconsistências decorrentes de falhas de reconhecimento, da qualidade da imagem ou das características gráficas do

documento analisado. Por essa razão, após a etapa de extração descrita no item 5.10, a solução executa o pós-processamento dos campos identificados, com a finalidade de limpar, padronizar e refinar os valores obtidos antes de sua apresentação ao usuário para validação.

O pós-processamento consiste na aplicação de procedimentos de limpeza, normalização e refinamento aos campos previamente extraídos, com o objetivo de corrigir inconsistências e preparar os valores identificados para as etapas posteriores de validação, conferência e armazenamento. Nessa etapa, foram implementadas as seguintes operações:

a) remoção de caracteres especiais indesejados presentes nos valores extraídos;
b) correção de espaços excedentes e normalização da formatação textual;
c) padronização de letras maiúsculas e minúsculas quando necessário;
d) eliminação de símbolos e caracteres decorrentes de falhas do reconhecimento óptico;
e) substituição de caracteres frequentemente confundidos pelo OCR, como “O” por “0”, “I” por “1” e “S” por “5”, quando essa substituição for compatível com o tipo de campo analisado;
f) reconstrução e normalização de datas extraídas de forma incompleta ou com caracteres incompatíveis;
g) filtragem de informações incompatíveis com os padrões definidos para cada campo;
h) comparação dos valores extraídos a partir das diferentes configurações de segmentação de página utilizadas pelo Tesseract;
i) seleção do valor considerado mais adequado para cada campo documental.

Essa etapa é necessária porque um mesmo campo pode ser reconhecido de formas diferentes conforme o modo de segmentação utilizado pelo Tesseract. Assim, a comparação entre os valores extraídos a partir dos diferentes PSMs foi utilizada para selecionar o resultado considerado mais adequado e encaminhá-lo às etapas seguintes de validação, conferência e armazenamento.

A adoção de mecanismos de pós-processamento complementa as etapas de reconhecimento e extração, pois permite corrigir, padronizar e refinar os valores identificados antes de sua validação final pelo usuário. Dessa forma, a solução não se limita à conversão da imagem em texto e à localização automática dos campos, mas

também realiza tratamentos sobre os dados extraídos para melhorar sua apresentação e facilitar a conferência.

5.12 Validação, rastreabilidade e armazenamento

Após o reconhecimento óptico de caracteres, a extração e o pós-processamento, os campos identificados são submetidos a regras de validação compatíveis com o tipo documental e, posteriormente, apresentados ao usuário para conferência, correção e confirmação.

Essa etapa foi concebida com o objetivo de garantir que as informações extraídas dos documentos não sejam utilizadas de forma automática e irrestrita, mas passem por um processo de conferência que permita ao operador confirmar sua correção antes do registro definitivo.

A solução desenvolvida adota uma estratégia de validação humana integrada ao ciclo de processamento, na qual o sistema atua como ferramenta de apoio ao servidor responsável pela matrícula, automatizando etapas iniciais de extração e organização dos dados, sem dispensar a conferência final pelo usuário. A decisão final sobre a validade das informações permanece sob responsabilidade do usuário.

Após o processamento das informações, os dados extraídos são apresentados ao usuário em telas de validação da aplicação. Compete ao usuário analisar os campos identificados automaticamente, realizar correções quando necessário e confirmar as informações antes da exportação e do armazenamento dos resultados.

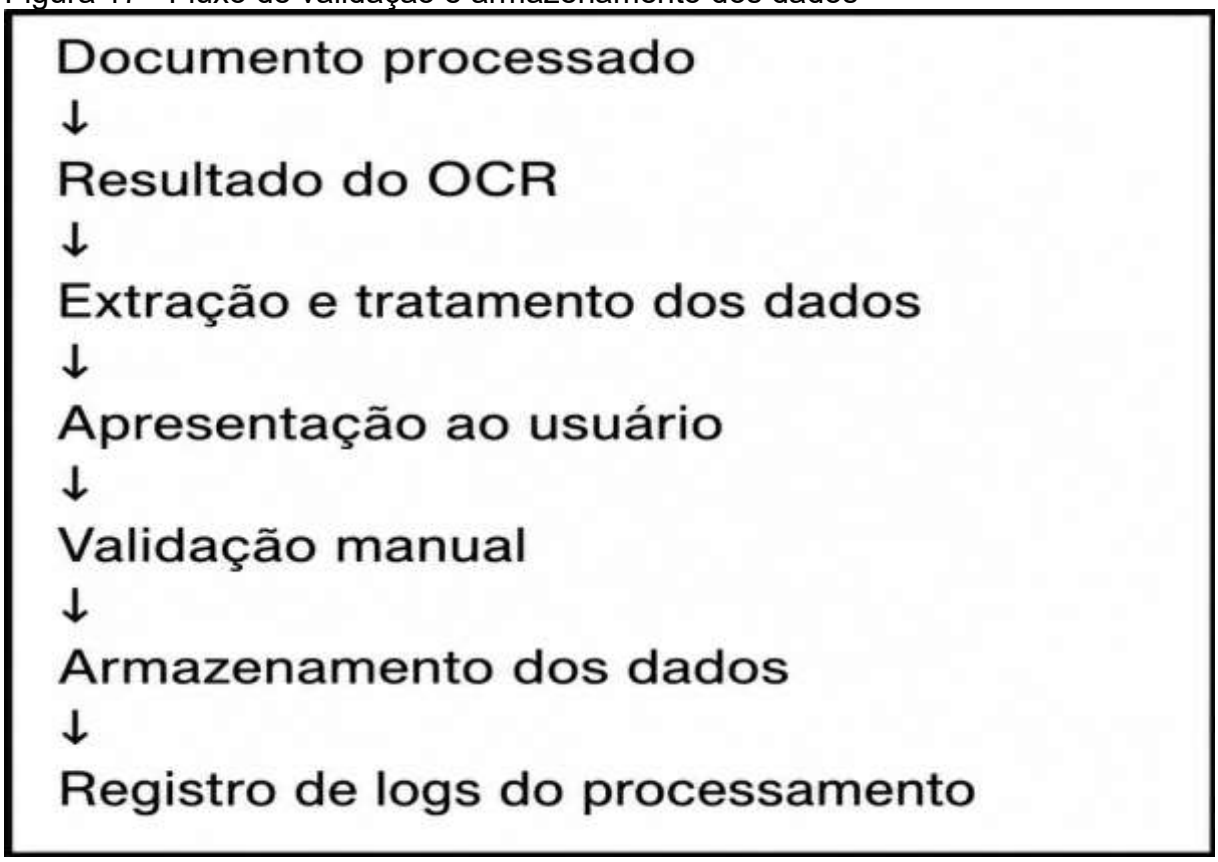
Após essa conferência, os dados podem ser armazenados em arquivo no formato CSV, permitindo sua utilização posterior no fluxo de matrícula dos estudantes da SRA.

A solução também registra eventos técnicos relacionados às operações executadas, incluindo data e hora, tipo documental, etapa do processamento, configurações empregadas, status da execução, alertas e mensagens de erro. Os valores pessoais extraídos dos documentos não são incluídos nesses registros.

Esses registros permitem acompanhar o funcionamento da aplicação e facilitar a rastreabilidade das operações realizadas, sem substituir mecanismos formais de auditoria institucional.

O fluxo da solução pode ser representado conforme apresentado na Figura 17.

Figura 17 - Fluxo de validação e armazenamento dos dados



Fonte: Elaborado pelo autor (2026).

A implementação dessa camada final de controle demonstra que o objetivo da solução não é substituir a atuação dos servidores da SRA, mas oferecer apoio às etapas iniciais de extração, organização e conferência das informações. Eventuais efeitos sobre tempo, incidência de erros ou eficiência operacional não foram mensurados nesta pesquisa.

5.12.1 Validação assistida pelo usuário

Após a escolha do tipo documental na tela inicial, a aplicação direciona o usuário para uma tela de carregamento e validação. Foram desenvolvidas duas opções de tela: uma destinada aos documentos do tipo CNH e outra destinada aos documentos do tipo RG ou CIN.

Nessas telas, o usuário realiza o carregamento do arquivo correspondente ao documento selecionado, aciona o fluxo de processamento e, após a execução das etapas automáticas de pré-processamento, OCR, extração e pós-processamento textual, visualiza os campos preenchidos automaticamente pela solução.

Os dados obtidos não são considerados definitivos. Antes do armazenamento, as informações passam por validação assistida pelo usuário, na qual os campos extraídos são apresentados em campos editáveis para conferência, correção e confirmação.

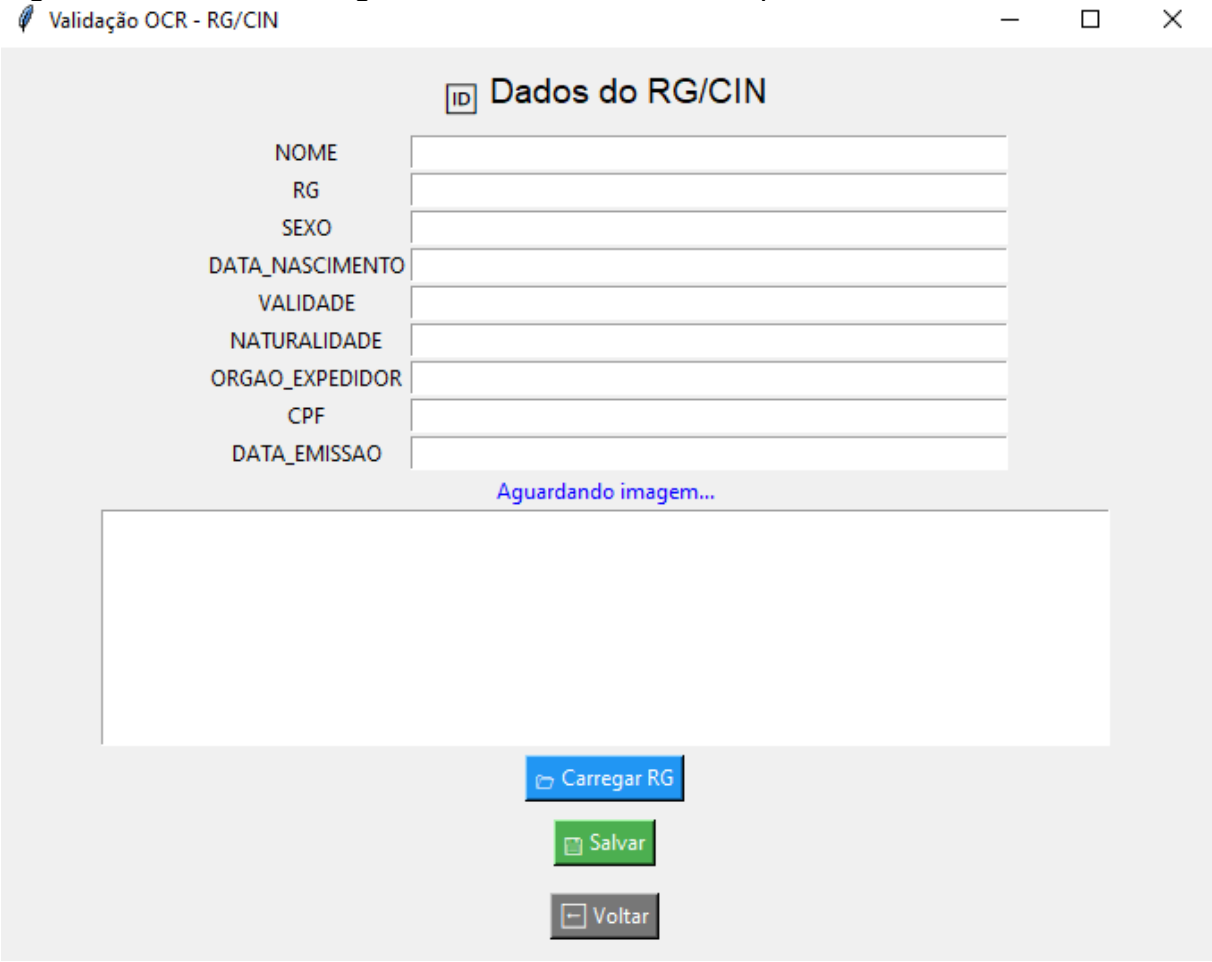
Essa estratégia foi adotada em razão das limitações inerentes ao OCR, especialmente em documentos com variações de qualidade, iluminação, resolução, enquadramento e organização visual. Dessa forma, a solução não substitui a atuação do servidor responsável pela matrícula, mas oferece apoio ao processo de extração inicial dos dados, mantendo a decisão final sobre a validade das informações sob responsabilidade do usuário.

As Figuras 18 e 19 apresentam as telas de carregamento e validação correspondentes, respectivamente, aos documentos do tipo CNH e RG/CIN, antes da seleção e do processamento da imagem.

Figura 18 - Tela de carregamento de documento do tipo CNH

Fonte: Elaborado pelo autor (2026).

Figura 19 - Tela de carregamento de documento do tipo RG/CIN



Fonte: Elaborado pelo autor (2026).

Após o carregamento e o processamento do documento, os dados identificados automaticamente são exibidos em campos editáveis, permitindo que o usuário compare as informações apresentadas com o documento original. Caso identifique algum dado incorreto, incompleto ou incompatível, o usuário pode realizar a correção manual antes da confirmação e do armazenamento dos resultados.

Dessa forma, a validação assistida pelo usuário atua como uma camada de controle entre a extração automática e a exportação dos dados, reduzindo o risco de utilização de informações incorretas e reforçando o caráter de apoio da solução proposta.

5.12.2 Armazenamento dos dados em arquivo CSV

Após a validação das informações pelo usuário, os dados podem ser armazenados em arquivo no formato CSV. No contexto da SRA, em que as

informações dos estudantes são posteriormente inseridas no SUAP, a geração de um arquivo CSV representa uma forma de organizar os dados extraídos em estrutura tabular, facilitando sua conferência e eventual utilização em rotinas administrativas de exportação, importação ou integração com sistemas institucionais.

Embora esta pesquisa não tenha realizado testes diretamente com os servidores da SRA nem implementado integração automática com o SUAP, a exportação em CSV foi adotada por ser compatível com procedimentos comuns de organização e transferência de dados entre ferramentas auxiliares e sistemas acadêmicos.

A escolha do formato CSV ocorreu em razão de suas características de simplicidade, portabilidade e ampla compatibilidade com diferentes ferramentas computacionais, incluindo softwares de planilhas eletrônicas e sistemas de gerenciamento de dados. Além disso, esse formato permite que os resultados obtidos pelo sistema sejam armazenados de maneira estruturada, sem dependência de um banco de dados específico ou de uma plataforma externa.

No formato adotado, cada linha do arquivo CSV corresponde a um documento processado, enquanto as colunas representam o tipo documental e os campos extraídos ou validados pela solução: nome, número do RG ou da identidade, sexo, data de nascimento, data de validade, naturalidade, órgão expedidor, CPF, data de emissão, categoria, nome do pai, nome da mãe e número de registro, conforme o documento analisado. Os campos que não se aplicam a determinado tipo documental permanecem vazios, permitindo organizar RG, CIN e CNH em uma estrutura tabular comum.

A Figura 20 apresenta um recorte esquemático da estrutura do arquivo CSV, com algumas das colunas geradas pela solução e dados exclusivamente fictícios.

Figura 20 – Recorte esquemático da estrutura do arquivo CSV gerado pela solução

tipo_documento	nome	cpf	data_nascimento	identidade	categoria	validade
CNH	PESSOA TESTE	000.000.000-00	01/01/2000	000000000	B	01/01/2030

Fonte: Elaborado pelo autor (2026).

Nesse procedimento, o arquivo CSV é gravado em modo de inclusão, permitindo que novos registros sejam adicionados sem sobrescrever os dados anteriormente armazenados.

5.12.3 Registro de execução e rastreabilidade técnica

Além do armazenamento dos dados extraídos, a solução desenvolvida implementa um sistema de registros de eventos (logs) com a finalidade de acompanhar as principais etapas do processamento realizado. Esses registros permitem documentar a execução da aplicação, identificar falhas técnicas e verificar em qual etapa eventual inconsistência ocorreu.

Os logs foram estruturados para registrar informações técnicas do funcionamento do sistema, sem armazenar desnecessariamente dados pessoais extraídos dos documentos. Dessa forma, informações como nome, CPF, filiação, número de documento e data de nascimento não são registradas nos logs, reduzindo o risco de exposição indevida dos dados processados.

Na estrutura adotada, existem informações essenciais que aparecem em todos os registros de log e informações variáveis, que são registradas apenas quando determinada situação ocorre durante a execução.

As informações essenciais são aquelas necessárias para identificar cronologicamente o evento, localizar a etapa do processamento e compreender o estado da execução. Assim, cada registro de log contém:

a) data e hora utilizados para indicar o momento exato em que determinada ação ocorreu;
b) etapa do processamento, como carregamento da imagem, pré-processamento, OCR, extração, validação, conferência ou exportação;
c) status da etapa, indicando se a operação foi iniciada, concluída, apresentou alerta ou resultou em erro;
d) mensagem técnica, descrevendo de forma objetiva o que ocorreu durante aquela etapa;

Além desses elementos fixos, alguns registros podem variar conforme o tipo de documento processado, a qualidade da imagem e o resultado do OCR. Essas informações variáveis não aparecem em todos os logs, pois dependem das condições específicas de cada execução. Entre elas destacam-se:

a) tipo de documento identificado, como RG, CIN ou CNH, pois cada modelo documental possui campos e regras de extração diferentes;
b) modos PSM utilizados pelo Tesseract OCR, especialmente PSM 4, PSM 6 e PSM 11, quando a execução envolver a comparação entre diferentes

formas de segmentação textual;

c) ocorrência de campo não localizado, registrada quando determinada informação esperada não é identificada no documento processado;
d) ocorrência de inconsistência de validação, registrada quando algum campo extraído não atende ao padrão esperado, como datas incompletas, CPF em formato inválido ou ausência de campo obrigatório;
e) mensagens de erro, registradas apenas quando ocorre falha no carregamento do arquivo, incompatibilidade de formato, erro de leitura da imagem ou interrupção de alguma etapa do fluxo.

Essa diferenciação entre informações essenciais e variáveis é necessária porque nem todos os documentos passam pelas mesmas situações durante o processamento. Um documento com boa qualidade de imagem pode gerar apenas registros informativos de execução concluída, enquanto uma imagem com baixa resolução, sombras ou campos não reconhecidos pode gerar alertas ou erros específicos. Da mesma forma, documentos diferentes exigem campos diferentes: a CNH possui dados como categoria e número de registro, enquanto RG e CIN apresentam campos próprios de identificação civil.

A Figura 21 apresenta um exemplo simplificado de registro de log gerado pela aplicação.

Figura 21 - Exemplo simplificado de registro de log

```
Data/Hora: 15/07/2026 09:44:23
================================================
[09:44:23] INICIANDO PROCESSAMENTO DA IMAGEM
[09:44:23] === OCR TESSERACT ===
[09:44:23] Configuração 1: --psm 6 --oem 3
[09:44:36] Texto extraído (config 1)
------------------------------------
[09:44:36] Configuração 2: --psm 4 --oem 3
[09:44:50] Texto extraído (config 2)
------------------------------------
[09:44:50] Configuração 3: --psm 11 --oem 3
[09:44:59] Texto extraído (config 3)
------------------------------------
[09:44:59] === TEXTO SELECIONADO (MAIS COMPLETO) ===
[09:44:59] === EXTRAINDO TODOS OS DADOS ===
[09:44:59] === BUSCA NOME ===
[09:44:59] Campo nome localizado
[09:44:59] === ORGANIZAÇÃO DE DATAS ===
[09:44:59] Campo data de nascimento localizado
[09:44:59] === BUSCA CPF ===
[09:44:59] Campo CPF localizado
[09:44:59] === BUSCA IDENTIDADE ===
[09:44:59] Campo identidade localizado
[09:44:59] === BUSCA CATEGORIA ===
[09:44:59] Campo categoria localizado
[09:44:59] === BUSCA REGISTRO CNH ===
[09:44:59] Campo número de registro localizado
[09:44:59] === DADOS FINAIS EXTRAÍDOS ===
```

Fonte: Elaborado pelo autor (2026)

Portanto, os logs não têm a função de armazenar os dados pessoais extraídos dos documentos, mas sim de registrar tecnicamente o comportamento da aplicação durante o processamento. Esse mecanismo contribui para a rastreabilidade da execução, para a identificação de falhas e para a manutenção futura da solução.

5.13 Síntese do modelo computacional

O modelo computacional desenvolvido foi estruturado como um fluxo integrado de processamento, composto pelas etapas de aquisição da imagem, pré-processamento, reconhecimento óptico de caracteres, extração de informações, pós-processamento textual, conferência pelo usuário, armazenamento dos resultados e registro de eventos.

A integração entre técnicas de Processamento Digital de Imagens, o mecanismo Tesseract OCR, expressões regulares, telas de validação assistida pelo usuário, exportação em formato CSV e geração de logs permitiu a construção de um protótipo capaz de apoiar a extração, organização e conferência de informações documentais relacionadas ao processo de matrícula acadêmica.

As telas de validação, por meio dos campos editáveis, permitem que o usuário revise os dados extraídos e realize correções manuais antes do armazenamento. Dessa forma, a solução não elimina a atuação humana, mas disponibiliza um fluxo de apoio à conferência das informações obtidas a partir dos documentos analisados.

O armazenamento em arquivo CSV organiza os dados extraídos em formato tabular, possibilitando sua análise posterior e eventual utilização em rotinas administrativas. Já os logs do sistema registram eventos técnicos do processamento, como etapas executadas, status e falhas, funcionando como mecanismo de rastreabilidade técnica da execução da aplicação.

Dessa forma, o modelo apresentado nesta seção não deve ser compreendido como uma solução validada em operação real pela SRA, mas como um protótipo desenvolvido e testado em ambiente laboratorial controlado. A etapa seguinte do trabalho apresenta os resultados experimentais obtidos com documentos de identificação, permitindo avaliar quantitativamente o desempenho e as limitações da solução proposta.

6 RESULTADOS E DISCUSSÃO

Após a conclusão do desenvolvimento da solução proposta, foram realizados testes em ambiente laboratorial controlado, sem integração ao SUAP e sem validação operacional com os servidores da SRA, com o objetivo de avaliar o desempenho técnico do modelo computacional na extração de informações presentes em documentos oficiais utilizados nos processos de matrícula da SRA do IF Baiano – Campus Catu.

Para a composição da base experimental, foram utilizados 30 documentos, distribuídos entre 10 RG, 10 CIN e 10 CNH, contemplando imagens obtidas por diferentes meios, incluindo fotografias realizadas por dispositivos móveis, arquivos digitais e documentos sintéticos gerados por ferramentas de Inteligência Artificial. A utilização de documentos sintéticos teve como finalidade ampliar a base de testes e preservar a privacidade dos dados pessoais, em conformidade com a LGPD (Lei nº 13.709/2018).

Essa estratégia permitiu avaliar o comportamento da solução diante de diferentes condições de resolução, enquadramento, iluminação e qualidade visual das imagens utilizadas nos experimentos. Ressalta-se, contudo, que essas condições representam variações controladas de teste e não equivalem à validação do sistema em ambiente real de operação da SRA.

Para avaliar o comportamento da implementação em condições variadas de captura, foram aplicadas intencionalmente algumas alterações nas imagens das amostras, como diferentes níveis de iluminação, ruído digital, baixa resolução e variações de contraste. Essas alterações buscaram representar situações que podem ocorrer durante a captura de documentos, sem caracterizar teste operacional em ambiente institucional.

Conforme os campos definidos no Quadro 5, foram avaliados sete campos em cada RG, oito campos em cada CIN e dez campos em cada CNH. Considerando as dez amostras de cada tipo documental, foram analisadas 70 ocorrências de campos no RG, 80 na CIN e 100 na CNH, totalizando 250 ocorrências de campos documentais.

Para a avaliação do desempenho da solução proposta, foram utilizados os critérios de classificação apresentados no Quadro 6, estabelecidos na etapa metodológica de verificação do ciclo PDCA. Dessa forma, os resultados obtidos pela implementação foram classificados em completos, parciais e falhos.

O Quadro 7 resume os resultados encontrados durante os testes.

Quadro 7 - Resultados dos testes de extração automática

Documento	Ocorrências	Completos	Parciais	Falhos	Taxa de Acerto - TA
RG	70	27	26	17	38,6%
CIN	80	54	9	17	67,5%
CNH	100	70	10	20	70,0%
Total	250	151	45	54	60,4%

Fonte: Elaborado pelo autor (2026)

Os resultados apresentados no Quadro 7 demonstram que o desempenho da solução variou conforme o tipo documental analisado. No RG, 27 das 70 ocorrências avaliadas foram classificadas como completas, resultando em Taxa de Acerto de 38,6%. Além disso, 26 resultados foram classificados como parciais e 17 como falhos.

Na CIN, 54 das 80 ocorrências avaliadas foram classificadas como completas, nove foram classificados como parciais e 17 como falhos, resultando em Taxa de Acerto de 67,5%. Na CNH, foram identificados 70 resultados completos, dez parciais e vinte falhos, entre cem ocorrências avaliadas, correspondendo a uma Taxa de Acerto de 70,0%.

Em termos de reconhecimento integral, a CNH apresentou o maior percentual, seguida pela CIN e pelo RG. Uma possível explicação para essa diferença está na maior uniformidade visual observada nas amostras de CIN e CNH, em comparação com a diversidade de layouts dos RGs analisados. Entretanto, como a pesquisa não isolou quantitativamente o efeito do layout, do estado de conservação ou da qualidade da imagem, não é possível atribuir causalidade individual a esses fatores.

Conforme os critérios estabelecidos na metodologia, a Taxa de Acerto considera exclusivamente os campos classificados como completos e representa o indicador principal de reconhecimento integral. Os campos parciais não foram contabilizados como acertos, pois ainda exigem conferência e correção manual.

Entretanto, como a solução foi desenvolvida para operar com validação assistida pelo usuário, os resultados parciais foram analisados separadamente. Embora não possam ser utilizados diretamente sem revisão, esses resultados conservaram conteúdo suficiente para permitir a identificação do dado e apoiar sua correção. Por essa razão, calculou-se, de forma complementar, a Taxa de Reconhecimento Aproveitável, reunindo os campos completos e parciais. Os resultados são apresentados no Quadro 8.

Quadro 8 - Taxa de Reconhecimento Aproveitável por tipo documental

Documento	Ocorrências avaliadas	Completos ou parciais	Taxa de Reconhecimento Aproveitável - TRA
RG	70	53	75,7%
CIN	80	63	78,8%
CNH	100	80	80,0%
Total	250	196	78,4%

Fonte: Elaborado pelo autor (2026)

A Taxa de Reconhecimento Aproveitável foi de 75,7% para o RG, 78,8% para a CIN e 80,0% para a CNH. Considerando conjuntamente os três tipos documentais, a TRA geral foi de 78,4%.

Esses percentuais não significam que todos os campos foram reconhecidos integralmente. A TRA reúne resultados completos e parciais e, portanto, inclui campos que ainda necessitaram de conferência e correção manual. Sua interpretação deve ocorrer em conjunto com a Taxa de Acerto.

Essa distinção é especialmente relevante no RG. Embora sua TRA tenha alcançado 75,7%, somente 38,6% das ocorrências foram reconhecidas integralmente. Assim, uma parcela significativa do reconhecimento aproveitável do RG decorreu de resultados parciais. Na CNH, por outro lado, a TRA de 80,0% foi composta predominantemente por resultados completos, que corresponderam a 70,0% das ocorrências avaliadas.

Embora Barbosa (2017) também tenha empregado PDI e Tesseract OCR, os percentuais obtidos não são diretamente comparáveis aos desta pesquisa, devido às diferenças entre os objetos analisados, as bases experimentais, as unidades de análise e os critérios de classificação adotados. A aproximação entre os estudos limita-se à constatação de que a qualidade das imagens e as etapas de pré-processamento influenciam o desempenho do reconhecimento textual.

Observa-se, ainda, que documentos com maior padronização visual, incluindo a CIN e a CNH, apresentaram melhor desempenho quando comparados ao RG. Esse resultado também se relaciona às observações de Uezima et al. (2018), segundo as quais a qualidade da imagem e a legibilidade dos caracteres influenciam diretamente o desempenho de aplicações baseadas em OCR.

Dessa forma, os resultados obtidos indicam que a utilização combinada de técnicas de pré-processamento, OCR, regras documentais e conferência pelo usuário compõe um fluxo de tratamento, validação e revisão dos dados extraídos, sem

dispensar a atuação do operador responsável pela conferência final das informações. A análise dos registros de log e dos resultados classificados permitiu identificar alguns fatores associados aos erros parciais e às falhas observadas durante os testes.

Entre eles, destacam-se:

a) Na solução testada, foram observadas confusões entre caracteres visualmente semelhantes, incluindo “2” e “3”, “O” e “0”, bem como “I” e “1”, especialmente em imagens com baixa qualidade visual. Essas inconsistências não devem ser atribuídas isoladamente ao mecanismo OCR utilizado, pois podem decorrer da combinação entre qualidade da imagem de entrada, etapas de pré-processamento, configuração dos modos de segmentação, reconhecimento textual e regras de extração adotadas na implementação;
b) Critérios de extração baseados em expressões regulares: as regras de extração foram definidas de forma restritiva para evitar a aceitação de padrões inválidos. Por esse motivo, quando o texto resultante do processamento não apresentava separadores, como pontos, barras ou hífens no formato esperado, a identificação automática do campo podia ser comprometida;
c) Variação de layout entre documentos: a CIN e a CNH apresentam maior padronização visual, enquanto o RG possui maior diversidade de formatos, elementos gráficos de segurança e variações de conservação, fatores que dificultaram o reconhecimento e a extração automática;
d) Qualidade da imagem de entrada: iluminação inadequada, baixa resolução, sombras, ruídos e contraste insuficiente interferiram no desempenho do fluxo de processamento.

As etapas de pré-processamento contribuíram para preparar as imagens antes da etapa de reconhecimento textual, especialmente por meio da conversão para tons de cinza, binarização e aplicação de filtro de mediana. Entretanto, os resultados demonstram que essas etapas não eliminam completamente os problemas decorrentes de baixa qualidade da imagem, variações de layout ou elementos gráficos presentes nos documentos.

Outra estratégia implementada foi a utilização combinada dos modos PSM 4, 6 e 11 do Tesseract OCR. Essa abordagem permitiu comparar diferentes formas de segmentação textual no processamento dos documentos, uma vez que determinadas

informações poderiam ser reconhecidas em um modo de segmentação e não em outro.

Os resultados indicam potencial de apoio à extração e à conferência inicial de dados documentais. Entretanto, como os experimentos foram realizados exclusivamente em ambiente laboratorial controlado, sem participação dos servidores, integração ao SUAP, medição de tempo, esforço de correção ou comparação com o procedimento manual, não é possível afirmar que a solução esteja pronta para uso institucional ou que reduza o tempo e o esforço de cadastramento. Essas dimensões dependem de validação futura no ambiente real da SRA.

7 CONSIDERAÇÕES FINAIS

Este trabalho teve como objetivo desenvolver um modelo computacional capaz de auxiliar o processo de matrícula e cadastramento de estudantes na SRA do IF Baiano – Campus Catu, por meio da captura, identificação, validação e organização automática de informações presentes em documentos oficiais, como RG, CIN e CNH.

A pesquisa partiu da necessidade de apoiar uma atividade administrativa que ainda depende da consulta e digitação manual de informações documentais no SUAP. Nesse contexto, a solução proposta não foi concebida para substituir a atuação dos servidores, mas para oferecer um fluxo de apoio à extração inicial, conferência e organização das informações presentes nos documentos analisados.

Em relação aos objetivos específicos, o primeiro objetivo, referente ao estudo de técnicas de PDI aplicáveis à extração automática de informações em RG, CIN e CNH, foi contemplado na fundamentação teórica, especialmente na discussão sobre PDI, OCR, limitações do reconhecimento textual e aplicações em documentos oficiais.

O segundo objetivo, relacionado à aplicação de técnicas de pré-processamento para melhoria da qualidade das imagens utilizadas pelo OCR, foi contemplado na implementação das etapas de conversão para tons de cinza, binarização pelo método de Otsu e redução de ruídos por filtro de mediana.

O terceiro objetivo, referente à implementação de um modelo de reconhecimento e extração automática de dados utilizando OCR, foi contemplado na construção do modelo computacional responsável pelo carregamento da imagem, aplicação do reconhecimento óptico de caracteres e extração de campos documentais específicos.

O quarto objetivo, relacionado à investigação de mecanismos de validação das informações extraídas, foi contemplado por meio da utilização de expressões regulares, regras documentais, pós-processamento textual e telas de validação assistida pelo usuário, permitindo ao usuário revisar e corrigir os dados antes do armazenamento.

O quinto objetivo, referente à avaliação do desempenho da solução em ambiente laboratorial controlado, foi contemplado na Seção 6, por meio da classificação dos campos extraídos como completos, parciais ou falhos e do cálculo dos indicadores Taxa de Acerto (TA) e Taxa de Reconhecimento Aproveitável (TRA).

À luz dos resultados obtidos e dos critérios definidos nesta pesquisa, considera-

se que a hipótese foi parcialmente confirmada. A integração entre PDI, OCR, regras documentais e validação assistida permitiu reconhecer e organizar, de forma integral ou parcialmente aproveitável, parte dos campos analisados. Entretanto, a hipótese não pode ser considerada integralmente confirmada quanto ao apoio operacional ao processo de matrícula, pois a solução apresentou resultados dependentes de correção manual e não foi validada em ambiente real da SRA, com participação dos servidores ou comparação com o procedimento manual.

Em resposta à questão de pesquisa, o modelo computacional demonstrou, em ambiente laboratorial controlado, capacidade de apoiar a extração inicial, a organização e a conferência de parte dos campos presentes em RG, CIN e CNH. Entretanto, não foi possível determinar a extensão desse apoio no processo real de matrícula, pois não houve testes com os servidores da SRA, medição de tempo, avaliação do esforço de correção ou comparação com o procedimento manual.

Os resultados obtidos indicaram que o desempenho da solução variou conforme o tipo de documento analisado. A CIN e a CNH apresentaram melhores resultados em comparação ao RG, possivelmente em razão da maior padronização visual e da organização mais uniforme dos campos. O RG, por sua vez, apresentou maior incidência de resultados parciais e falhos, especialmente em razão da diversidade de layouts, dos elementos gráficos de segurança e das variações na qualidade visual dos documentos.

Os resultados demonstraram TA de 38,6% para o RG, 67,5% para a CIN e 70,0% para a CNH. Como indicador complementar, a TRA foi de 75,7%, 78,8% e 80,0%, respectivamente. Considerando as 250 ocorrências de campos documentais, a TA geral foi de 60,4% e a TRA geral foi de 78,4%.

Esses resultados demonstram que o desempenho variou conforme o tipo documental e que parte dos campos classificados como parcialmente reconhecidos permaneceu dependente de intervenção humana. Consequentemente, os percentuais de TRA não devem ser interpretados como reconhecimento integral nem como comprovação de redução de tempo ou de esforço operacional.

Durante os testes, também foram identificadas limitações relevantes. As falhas e inconsistências observadas não devem ser atribuídas isoladamente ao mecanismo OCR utilizado, pois decorrem do conjunto da solução proposta, incluindo qualidade da imagem de entrada, pré-processamento aplicado, configuração dos modos de segmentação, regras de extração, expressões regulares e variações de layout dos

documentos.

É importante destacar que a avaliação realizada nesta pesquisa ocorreu exclusivamente em ambiente laboratorial controlado. Não houve implantação da ferramenta em ambiente real da SRA, validação operacional com servidores, medição cronometrada do tempo de uso ou integração com o SUAP. Dessa forma, os resultados devem ser compreendidos como uma avaliação experimental do protótipo desenvolvido, e não como comprovação de desempenho em operação institucional.

Também constitui limitação metodológica a utilização das mesmas amostras nos ciclos de desenvolvimento, ajuste e avaliação final do protótipo. Como não foi utilizado um conjunto independente de teste, os resultados possuem caráter exploratório e podem refletir adaptação da implementação às características da base experimental.

Outra limitação da pesquisa está relacionada às métricas utilizadas. A avaliação foi realizada por meio da classificação categórica de cada campo como completo, parcial ou falho, acompanhada do cálculo da Taxa de Acerto e da Taxa de Reconhecimento Aproveitável. Essa abordagem permitiu avaliar o comportamento dos campos estruturados e sua possibilidade de utilização no fluxo de validação assistida, mas não mensurou a quantidade exata de caracteres ou palavras reconhecidos incorretamente, o tempo de processamento dos documentos, o esforço necessário para correção dos resultados parciais nem eventuais diferenças em relação ao procedimento manual de cadastramento.

A principal contribuição do trabalho está na sistematização e avaliação experimental de um modelo computacional aplicado a documentos oficiais brasileiros, evidenciando possibilidades e limitações da solução proposta em ambiente laboratorial controlado.

Como trabalhos futuros, recomenda-se ampliar a base experimental, separar conjuntos distintos para desenvolvimento e avaliação, realizar testes de campo com servidores da SRA e adotar métricas complementares de acurácia, erro por caractere e por palavra, tempo de processamento e esforço de correção. Também se sugere comparar outros mecanismos de OCR, ampliar os tipos documentais, validar o CPF pelos dígitos verificadores e avaliar eventual integração segura com o SUAP, observando autenticação, controle de acesso, rastreabilidade e proteção de dados pessoais.

REFERÊNCIAS

ALBUQUERQUE, Márcio Portes de; ALBUQUERQUE, Marcelo Portes de. **Processamento de imagens**: métodos e análises. Rio de Janeiro: Centro Brasileiro de Pesquisas Físicas, 2002. Disponível em: https://mesonpiold.cbpf.br/e2002/cursos/NotasAula/PDSI.pdf. Acesso em: 10 dez. 2025.

ÁVILA, T.; LANZA, B.; VALOTTO, D. **Transformação digital, tecnologia e inovação nos estados brasileiros**: os caminhos propostos para o período de 2023-2026. Curitiba: Red Académica de Gobierno Abierto, 2023.

BARBOSA, Alexandre Henrique C. **Processamento digital de imagens para o reconhecimento de placas de veículos**. 2017. 25 f. Trabalho de Conclusão de Curso (Bacharelado em Sistemas de Informação) – Unibalsas – Faculdade de Balsas, Balsas, 2017. Disponível em: https://www.unibalsas.edu.br/wp-content/uploads/2017/01/Alexandre.pdf. Acesso em: 5 ago. 2025.

BARELLI, Felipe. **Introdução à visão computacional**: uma abordagem prática com Python e OpenCV. São Paulo: Casa do Código, 2018.

BRASIL. **Decreto nº 10.278, de 18 de março de 2020**. Regulamenta o disposto no inciso X do caput do art. 3º da Lei nº 13.874, de 20 de setembro de 2019, e no art. 2º-A da Lei nº 12.682, de 9 de julho de 2012, para estabelecer a técnica e os requisitos para a digitalização de documentos públicos ou privados, a fim de que os documentos digitalizados produzam os mesmos efeitos legais dos documentos originais. Brasília, DF: Presidência da República, 2020. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2020/decreto/d10278.htm. Acesso em: 13 jul. 2026.

BRASIL. **Lei nº 8.159, de 8 de janeiro de 1991**. Dispõe sobre a política nacional de arquivos públicos e privados e dá outras providências. Brasília, DF: Presidência da República, 1991. Disponível em: https://www.planalto.gov.br/ccivil_03/leis/l8159.htm. Acesso em: 13 jul. 2026.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Presidência da República, 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 13 jul. 2026.

BRASIL. **Lei nº 14.129, de 29 de março de 2021**. Dispõe sobre princípios, regras e instrumentos para o Governo Digital e para o aumento da eficiência pública e altera a Lei nº 7.116, de 29 de agosto de 1983, a Lei nº 12.527, de 18 de novembro de 2011, a Lei nº 12.682, de 9 de julho de 2012, e a Lei nº 13.460, de 26 de junho de 2017. Brasília, DF: Presidência da República, 2021. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14129.htm. Acesso em: 13 jul. 2026.

CARVALHO, Emmanuel Aires Urquiza de. **Identificação óptica de caracteres de placas de trânsito**. 2015. 74 f. Trabalho de Conclusão de Curso (Bacharelado em Engenharia Elétrica) – Centro de Engenharia Elétrica e Informática, Universidade

Federal de Campina Grande, Campina Grande, 2015. Disponível em: https://dspace.sti.ufcg.edu.br/handle/riufcg/18510. Acesso em: 5 ago. 2025.

CASERTA, Thiago. **Transformação digital desmistificada**: como as revoluções digitais já mudaram o jogo, e como jogá-lo enquanto ainda é tempo. 1. ed. São Paulo: Gente Autoridade, 2023.

CONSELHO NACIONAL DE ARQUIVOS (Brasil). **e-ARQ Brasil**: modelo de requisitos para sistemas informatizados de gestão arquivística de documentos. Versão 2. Rio de Janeiro: Arquivo Nacional, 2022. Disponível em: https://www.gov.br/conarq/pt-br/centrais-de-conteudo/publicacoes/EARQV203MAI2022.pdf. Acesso em: 13 jul. 2026.

ESCOBAR, Fernando. Implementando a transformação digital. In: LOUREIRO, Geraldo (org.). **Reconstrução do Brasil pela transformação digital no setor público**. Brasília, DF: IBGP, 2020. p. 95-152. Disponível em: https://d1.awsstatic.com/WWPS/pdf/Livro_reconstrucao_do_brasil_pela_transformacao_digital_no_setor_publico.pdf. Acesso em: 1 ago. 2025.

GONZALEZ, R. C.; WOODS, R. E. **Digital image processing**. 4. ed. New York: Pearson Education, 2018.

GOOGLE. **Gemini**: sistema de inteligência artificial generativa. [S. l.]: Google, [s. d.]. Disponível em: https://gemini.google.com/. Acesso em: 1 ago. 2025.

GIL, Antonio Carlos. **Métodos e técnicas de pesquisa social.** 2. ed. São Paulo: Atlas, 1989.

LEITE, Guilherme Nascimento. **Implementação de algoritmo de processamento digital de imagens para segmentação de ovos de galinha**. 2024. 44 f. Trabalho de Conclusão de Curso (Graduação em Ciência da Computação) – Universidade Federal de Uberlândia, Uberlândia, 2024. Disponível em: https://repositorio.ufu.br/handle/123456789/43514. Acesso em: 3 ago. 2025.

MARQUES FILHO, Ogê; VIEIRA NETO, Hugo. **Processamento digital de imagens**. Rio de Janeiro: Brasport, 1999.

QUEIROZ, José Eustáquio Rangel de; GOMES, Herman Martins. Introdução ao processamento digital de imagens. RITA: **Revista de Informática Teórica e Aplicada**, Porto Alegre, v. 8, n. 1, p. 1-31, 2001. Disponível em: https://www.dsc.ufcg.edu.br/~hmg/disciplinas/graduacao/vc-2015.1/Rita-Tutorial-PDI.pdf. Acesso em: 11 dez. 2025.

REZENDE, André L. A.; JESUS, Cayo P. S. de; NOGUEIRA, Amanda E. S.; PEREIRA, Isis B. S.; S. NETO, Aderaldo C. da; COSTA, Gustavo J. da S.; SOUZA, Márcio V. S. Tecnologia Assistiva MINDER: possibilidades de um computador de baixo custo na inclusão sociodigital de sujeitos não videntes. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE), 32., 2021, Online. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2021. p. 586-597. DOI: https://doi.org/10.5753/sbie.2021.217539. Acesso em: 9 dez. 2025.

SCHALLMO, Daniel R. A.; WILLIAMS, Christopher A.; QUADROS, Ruy; FRANCO, Matheus M. V. **Transformação digital já!**: um guia para a digitalização do seu modelo de negócio. Brasília, DF: IEL/NC, 2021. Disponível em: https://static.portaldaindustria.com.br/media/filer_public/bf/8f/bf8f6144-df73-47d9-81d8-711777cfac14/livro_transformacao_digital_final_v2_2.pdf. Acesso em: 1 ago. 2025.

SMITH, Ray. An overview of the Tesseract OCR engine. In: INTERNATIONAL CONFERENCE ON DOCUMENT ANALYSIS AND RECOGNITION (ICDAR), 9., 2007, Curitiba. **Proceedings [...]**. Washington, DC: IEEE Computer Society, 2007. v. 2, p. 629-633. DOI: 10.1109/ICDAR.2007.4376991. Disponível em: https://research.google.com/pubs/archive/33418.pdf. Acesso em: 1 ago. 2025.

TESSERACT OCR. **Tesseract user manual**. [S. l.]: Tesseract OCR, [s. d.]. Disponível em: https://tesseract-ocr.github.io/tessdoc/. Acesso em: 1 ago. 2025.

THOMAS, Reginaldo. **Letras nas observações da CNH**: afinal, o que significa cada uma? UOL Carros, 27 jul. 2023. Atualizado em: 12 out. 2023. Disponível em: https://www.uol.com.br/carros/noticias/redacao/2023/07/27/letras-nas-observacoes-da-cnh.htm. Acesso em: 25 nov. 2025.

UEZIMA, Lucas K.; NISHIMOTO, Henrique Y. Y.; BARBOSA, Rafael; BASILE, Antonio Luiz. Utilizando o Tesseract OCR para reconhecimento de textos à mão. In: ESCOLA REGIONAL DE INFORMÁTICA DE SÃO PAULO (ERI-SP), 2018, São Paulo. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2018. Disponível em: https://dspace.mackenzie.br/items/b87a0d67-baf7-4837-8e9c-7f0e260c512c. Acesso em: 5 ago. 2025.

Documento Digitalizado Público

TCC - Versão final com ficha catalográfica

Assunto: TCC - Versão final com ficha catalográfica
Assinado por: Andre Rezende
Tipo do Documento: ANEXO
Situação: Finalizado
Nível de Acesso: Público
Tipo do Conferência: Cópia Simples

Documento assinado eletronicamente por:

- **Andre Luiz Andrade Rezende**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 24/07/2026 11:26:41.

Este documento foi armazenado no SUAP em 24/07/2026. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifbaiano.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 1367981
Código de Autenticação: 97a2287099