CEFET-MG
Centro Federal de Educação Tecnológica de Minas Gerais
Mestrado em Modelagem Matemática e Computacional

VISÃO COMPUTACIONAL APLICADA À SISTEMAS DE AUXÍLIO AO DIAGNÓSTICO: Identificação de pontos anatômicos no traçado cefalométrico

Germano Teixeira de Miranda

Orientador: Prof. Dr. Flávio Vinícius Cruzeiro Martins
CEFET-MG

Belo Horizonte
Dezembro de 2023

Germano Teixeira de Miranda

VISÃO COMPUTACIONAL APLICADA À SISTEMAS DE AUXÍLIO AO DIAGNÓSTICO: Identificação de pontos anatômicos no traçado cefalométrico

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Matemática e Computacional do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito parcial para a obtenção do título de Mestre em Modelagem Matemática e Computacional.

Área de concentração: Modelagem Matemática e Computacional.

Linha de pesquisa: Sistemas Inteligentes.

Orientador: Prof. Dr. Flávio Vinícius Cruzeiro Martins
CEFET-MG

Centro Federal de Educação Tecnológica de Minas Gerais
Mestrado em Modelagem Matemática e Computacional
Belo Horizonte
Dezembro de 2023

Página destinada à ficha catalográfica.

Germano Teixeira de Miranda

VISÃO COMPUTACIONAL APLICADA À SISTEMAS DE AUXÍLIO AO DIAGNÓSTICO: Identificação de pontos anatômicos no traçado cefalométrico

Dissertação apresentada ao Programa de Pós-graduação em Modelagem Matemática e Computacional do Centro Federal de Educação Tecnológica de Minas Gerais, como requisito parcial para a obtenção do título de Mestre em Modelagem Matemática e Computacional.

Trabalho aprovado. Belo Horizonte, 23 de dezembro de 2023:

Prof. Dr. Flávio Vinícius Cruzeiro Martins
CEFET-MG

Prof. Dr. Alisson Marques Da Silva
CEFET-MG

Prof. Dr. Leniana Santos Neves
UFMG

Belo Horizonte
Dezembro de 2023

AGRADECIMENTOS

Gostaria de dedicar este capítulo para expressar minha profunda gratidão a todas as pessoas que foram fundamentais para o sucesso deste projeto de mestrado.

Primeiramente, gostaria de agradecer aos meus pais pelo constante incentivo à minha educação e por terem me apoiado incondicionalmente em todas as etapas deste processo. Sem o suporte deles, eu não teria chegado até aqui.

Agradeço também ao professor José Benedito Limongi, que me apresentou ao mundo da radiologia odontológica e me ensinou a fazer um exame de traçado cefalométrico. Suas aulas e orientações foram fundamentais para o desenvolvimento deste projeto e para minha formação como profissional. Espero que esse trabalho honre o seu legado.

Gostaria de agradecer aos meus sócios do Cfaz.net por terem compreendido a importância deste projeto de mestrado. A compreensão de vocês em me permitir abdicar de algumas horas da empresa para me dedicar ao estudo foi essencial para que eu pudesse me concentrar totalmente neste projeto.

Agradeço também ao professor Flávio Cruzeiro por assumir o risco de me orientar mesmo sem me conhecer previamente. Sua orientação foi fundamental para o desenvolvimento deste projeto e para o meu crescimento profissional.

Por fim, agradeço à equipe do mestrado em modelagem matemática que trabalha com dedicação para manter o mestrado em andamento, mesmo em meio às dificuldades enfrentadas durante a pandemia. Sem a colaboração e o esforço deles, este projeto não poderia ter sido concluído.

A todos vocês, meus sinceros agradecimentos. Espero que este projeto possa contribuir de alguma forma para a ciência e para o desenvolvimento da área de radiologia odontológica e modelagem matemática.

A Idade da Pedra não acabou por falta de pedras.

RESUMO

Sistemas computacionais de auxílio ao diagnóstico são amplamente usados nos mais diversos setores da área da saúde a fim de facilitar o trabalho dos especialistas envolvidos. Esses sistemas computacionais auxiliam os profissionais de saúde diminuindo a chance de erro, melhorando a precisão do exame, além de diminuir tempo e o custo para elaboração do diagnóstico. Mais especificamente na odontologia, o exame de traçado cefalométrico consiste em fazer medidas lineares e angulares na face para embasar o diagnóstico e o tratamento ortodôntico. Para a elaboração desse exame, as medidas são feitas em uma radiografia lateral da cabeça chamada telerradiografia lateral. O diagnóstico de um traçado cefalométrico é obtido através da avaliação quantitativa das medidas feitas na telerradiografia em comparação com o padrão ideal definido pela literatura. Essa característica quantitativa e bem definida do exame o torna um candidato promissor para aplicação de técnicas de visão computacional que possibilitem a obtenção do diagnóstico de forma automática, sem a intervenção manual do especialista. O objetivo do meu trabalho de mestrado é avaliar as técnicas de visão computacional existentes atualmente e aplicá-las na detecção automática dos pontos anatômicos necessários para a elaboração de um exame de traçado cefalométrico. Para a aplicação das técnicas de visão computacional, foi extraída uma base de dados para o treinamento a partir de mais de um milhão de exames de traçados cefalométricos realizados na ferramenta de traçado computadorizado do Cfaz.net, um dos softwares de traçado cefalométrico mais usados na América Latina. Essa base de treinamento foi utilizada para treinar e avaliar as técnicas de visão computacional que melhor conseguem realizar a tarefa de encontrar pontos anatômicos em uma telerradiografia. No decorrer do trabalho, foram exploradas técnicas que conseguissem realizar o exame de traçado cefalométrico de maneira satisfatória, permitindo, ao final, afirmar qual das técnicas é melhor comparativamente. Deste modo, concluiu-se que as técnicas de visão computacional disponíveis atualmente no mercado são capazes de realizar um exame de traçado cefalométrico de maneira satisfatória sem a intervenção do especialista.

Palavras-Chave: Visão Computacional, Traçado cefalométrico, Telerradiografia.

ABSTRACT

Computational systems for diagnostic assistance are widely used in various sectors of the healthcare field to facilitate the work of involved specialists. These computer systems aid healthcare professionals by reducing the chance of errors, improving the accuracy of the examination, as well as reducing the time and cost for diagnostic elaboration. More specifically in dentistry, the cephalometric tracing examination consists of making linear and angular measurements on the face to support the diagnosis and orthodontic treatment. For the elaboration of this examination, measurements are made on a lateral head radiograph called lateral cephalogram. The diagnosis of a cephalometric tracing is obtained through the quantitative evaluation of the measurements made in the cephalography compared to the ideal standard defined by the literature. This quantitative and well-defined characteristic of the examination makes it a promising candidate for the application of computer vision techniques that allow the diagnosis to be obtained automatically, without manual intervention from the specialist. The objective of my master's thesis is to evaluate the current computer vision techniques and apply them in the automatic detection of anatomical points necessary for the elaboration of a cephalometric tracing examination. For the application of computer vision techniques, a training database was extracted from more than one million cephalometric tracing examinations performed on the computerized tracing tool of Cfaz.net, one of the most used cephalometric tracing software in Latin America. This training base was used to train and evaluate the computer vision techniques that best manage to perform the task of finding anatomical points in a cephalography. Throughout the work, techniques that could satisfactorily perform the cephalometric tracing examination were explored, allowing, in the end, to state which of the techniques is comparatively better. Thus, it was concluded that the current computer vision techniques available in the market are capable of satisfactorily performing a cephalometric tracing examination without the intervention of the specialist.

keywords: Computer Vision. Lateral Cephalogram. Teleradiography.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIAÇÕES E SIGLAS

CEFET	Centro Federal de Educação Tecnológica
MLP	Multilayer Perceptron
CNN	*Convolutional Neural Network*
R-CNN	*Region Based Convolutional Neural Network*
ViT	*Vision Transformer*
COCO	*Common Objects in Context Dataset*

SUMÁRIO

1 Introdução

Saúde é um direito humano fundamental reconhecido por todos os foros mundiais e em todas as sociedades, sendo um dos direitos garantidos pela Declaração Universal dos Direitos Humanos de 1948, em seu artigo 25, i (ONU, 1948).

Os procedimentos operacionais da área da saúde geram um grande volume de dados médicos armazenados, entre eles imagens, exames, diagnósticos e procedimentos de tratamento. Esses dados compõem o prontuário eletrônico do paciente.

Conforme artigo 8º da Resolução nº 1.821/2017 do Conselho Federal de Medicina, a preservação dos prontuários dos pacientes com seu consequente armazenamento deve ser de no mínimo 20 (vinte) anos (BRASIL, 2017). Dada a obrigatoriedade de armazenamento dos prontuários, esse grande volume de dados históricos se torna uma valiosa fonte de conhecimento, com o potencial de ser usada para o auxílio do diagnóstico médico.

Nesse contexto, ganha destaque o sistema computacional de auxílio ao diagnóstico (*Computer Aided Diagnosis* - CAD). Esse sistema deve ser entendido como qualquer sistema que fornece o resultado de análises quantitativas automatizadas como uma “segunda opinião” para a tomada de decisões diagnósticas (DOI K, 2007, p. 198-211). Estudos clínicos revelam que a utilização do sistema CAD aumenta de forma estatisticamente significativa (21%) a detecção por radiologistas do câncer de mama. Os resultados também mostraram que o uso de sistema CAD melhora estatisticamente a precisão dos radiologistas na distinção de nódulos, microcalcificações e assimetrias em mamografias, na análise do tamanho do coração, na detecção de embolia pulmonar, entre outros (CALAS, GUTFILEN, PEREIRA, 2012).

Desse modo, fica evidente a importância do desenvolvimento de ferramentas e técnicas computacionais para auxílio ao diagnóstico de imagens médicas. Indo ao encontro com o exposto até o momento, o estudo da *International Data Corporation* indica que o investimento em tecnologias no setor de saúde na América Latina vai atingir US $1.931 milhões até 2022, o que equivale a 10 bilhões de reais (IDC, 2020), o que mais uma vez ressalta a importância das técnicas computacionais aplicadas ao setor da saúde.

A odontologia é uma parte fundamental da saúde e tem papel significativo na qualidade de vida da população. Segundo o Ministério da Saúde, a odontologia esteve à

margem das políticas públicas de saúde. O acesso dos brasileiros à saúde bucal era extremamente difícil e limitado. Esta demora na procura pelo atendimento aliada aos poucos serviços odontológicos oferecidos faziam com que o principal tratamento oferecido pela rede pública fosse a extração dentária, perpetuando a visão da odontologia mutiladora e do cirurgião-dentista (BRASIL, 2003).

Para mudar esse cenário, o governo brasiliero lançou em 2003 a Política Nacional de Saúde Bucal (PNSB) - Brasil sorridente. Essa política consiste em uma série de medidas que visam a garantir ações de promoção, prevenção e recuperação da saúde bucal dos brasileiros, fundamental para a saúde geral e qualidade de vida da população. Atualmente o Brasil possui uma das melhores odontologias do mundo. Em 2017, o ranking mundial de universidades - *Rankings by Subject* - considerou 3 universidades brasileiras entre as 5 melhores do mundo (CWUR, 2017).

Vista a relevância da odontologia para a saúde bucal brasileira, uma das especialidades dessa área é a radiologia odontológica, que consiste em usar radiografias para examinar os dentes e a face de uma pessoa. Uma radiografia, por sua vez, é uma imagem obtida através da exposição de uma pessoa à radiação ionizante para coletar imagens de áreas internas do organismo (FREITAS, 2004). Essas radiografias são usadas desde o século XIX para a visualização de fraturas, tumores e outros males.

As radiografias são de extrema importância para o planejamento e acompanhamento de doenças bucais. Existem diversas utilizadas pela odontologia, entre elas: (i) as radiografias intraorais, que são imagens da parte interior da boca com objetivo de observar os dentes, o ligamento periodontal e o osso alveolar; (ii) a radiografia panorâmica, muito utilizada para avaliar a estrutura óssea da mandíbula e maxila e os dentes em uma visão mais ampla; (iii) a radiografia lateral da cabeça ou telerradiografia lateral, que é usada para observar a estrutura óssea da face, os dentes e a relação entre as estruturas anatômicas do crânio.

Vale dizer que os exames radiográficos são solicitados por praticamente todos especialistas na área da Odontologia e usados como ferramenta pelos especialistas para a elaboração de um laudo diagnóstico no qual serão apresentadas as observações pertinentes quanto à imagem.

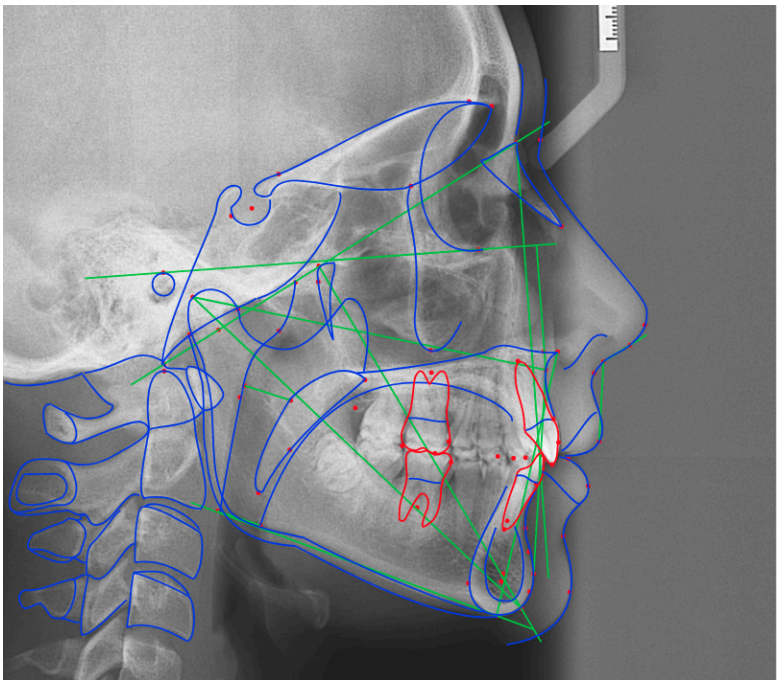
A radiografia lateral da cabeça, ou telerradiografia lateral, é utilizada para realização cefalograma. O traçado cefalométrico, por sua vez, consiste na marcação de

linhas e planos, traçados a partir de pontos cefalométricos previamente marcados. Por pontos cefalométricos, entende-se regiões precisamente especificadas por autores em uma radiografia lateral da face (telerradiografia lateral) na qual, além dos pontos, também são analisados planos, que são obtidos com base na marcação dos pontos (VILELLA, 2009).

A cefalometria ou análise cefalométrica é um exame utilizado para avaliar o crescimento e o desenvolvimento craniofacial de forma longitudinal, bem como auxiliar na elaboração do diagnóstico e plano de tratamento ortodôntico. Sua utilização ocorre nos casos de pacientes que utilizam aparelho ortopédico ou ortodôntico para corrigir desarmonias dentárias, faciais ou no sorriso, como as más oclusões de Classe II, Classe III, biprotrusão dentoalveolar, mordida aberta e mal posicionamento dos dentes em geral.

Para melhor entendimento quanto a criação de um cefalograma, a figura 1 mostra uma telerradiografia com pontos e planos demarcados.

Figura 1 - Desenho anatômico, pontos, linhas e planos cefalométricos



Fonte: Software de Cefalometria Computadorizada Cfaz.net, 2020

Ao resultado das mensurações lineares ou dos ângulos formados entre a linhas e planos, dá-se o nome de fatores cefalométricos. Esses fatores, juntamente com as

respectivas normas e desvios-padrão, são apresentados em uma tabela de resultados. A Figura 2 mostra parte de uma tabela de fatores proposta por Ricketts (VILELLA, 2009).

Figura 2 - Conjunto de fatores cefalométricos de Ricketts

#	Descrição	Valor	Padrão
Campo I - Problemas Dentários			
1	Relação Molar	-1.36 mm	-3.0 ± 3.0
2	Relação Canina	0.168 mm	-2.0 ± 3.0
3	Trespasse Horizontal do Incisivo (Overjet)	2.736 mm	2.5 ± 2.5
4	Trespasse Vertical do Incisivo (Overbite)	2.587 mm	2.5 ± 2.0
5	Extrusão do Incisivo Inferior	1.68 mm	1.25 ± 2.0
6	Ângulo Interincisivo	135.332 gr	130.0 ± 6.0

Fonte: Software de Cefalometria Computadorizada Cfaz.net, 2020

Ao conjunto das medidas lineares e angulares obtidas bem como suas interpretações, dá-se o nome de análise cefalométrica.

Os primeiros traçados cefalométricos eram feitos manualmente através da marcação de pontos na radiografia e posteriormente realizadas as medições dos fatores com transferidores de ângulos, réguas e esquadros. A partir da década de 60 os computadores começaram a ser utilizados para que o operador marcasse esses pontos na imagem da telerradiografia lateral e os observasse na tela do computador, permitindo assim que as medidas e cálculos pudessem ser computadas instantaneamente (PAIVA, 2008). A utilização de computadores contribuiu para a popularização da técnica cefalométrica, que hoje é largamente utilizada no mercado de odontologia.

Os avanços tecnológicos dos sistemas de auxílio ao diagnóstico, especialmente no campo da visão computacional, como já tratado acima, têm oferecido recursos e ferramentas para auxiliar especialistas na execução de exames de imagem. A visão computacional é um dos ramos da inteligência artificial que estuda o processamento de imagens do mundo real por um computador. Ela consiste em usar um computador para extração de informações significativas de uma imagem em formato digital. Essas informações permitem reconhecer, manipular e pensar sobre objetos que compõem uma imagem (BROWN; BALLARD, 1982).

Melhoramento de imagem, transformações, filtros, compressão de imagem, percepção de cores, estimativa da pose, extração de características e reconhecimento de objetos são alguns dos tópicos abordados pela visão computacional. Dessa maneira, essa

área da computação permite que computadores utilizem as informações fornecidas para executar tarefas inteligentes, simulando e aproximando-se da inteligência humana.

Aplicações de visão computacional podem ser encontradas nas mais diversas áreas do conhecimento. Como ilustração tem-se: o reconhecimento facial de imagens do rosto (LI, 2011), a segmentação de imagens para interpretação de um cubo de rubik (LETO, 2015), e a detecção de anomalias em radiografias (QUEK et al., 2003).

As técnicas de visão computacional são relativamente novas e sua utilização ainda não foi explorada em todo o universo de problemas aplicáveis. A odontologia, devido ao grande volume de imagens utilizadas, oferece muitas oportunidades para aplicação de técnicas de visão computacional. Sendo as radiografias ferramentas necessárias ao planejamento e acompanhamento de doenças bucais, tem-se na telerradiografia lateral a possibilidade de exploração quanto a aplicabilidade da visão computacional ao traçado cefalométrico. Esse, em especial, é um exame com algoritmo muito bem estruturado e com resultados quantitativos muito bem definidos, condição que permite explorar a aplicação de técnicas de visão computacional, sendo essa a proposta da presente pesquisa de mestrado.

1.1 Motivação

A última década apresentou resultados impressionantes em pesquisas relacionadas à visão computacional, tanto na melhora da performance quanto na confiabilidade dos resultados. Esses resultados foram obtidos pela adoção de técnicas de *machine learning*, mais especificamente redes neurais, e a grande quantidade de dados do mundo real disponíveis (SZELISKI, 2022).

Hoje existe um grande volume de radiografias armazenadas com exames de pacientes. Isso acontece por força de lei, já que as radiografias fazem parte do prontuário do paciente e a legislação brasileira que dispõe sobre a digitalização e a utilização de sistemas informatizados para a guarda prevê que o armazenamento e o manuseio de prontuário de paciente ocorram por pelo menos 20 anos. Veja-se artigo 6º da Lei 13.787/2018: “Decorrido o prazo mínimo de 20 (vinte) anos a partir do último registro, os prontuários em suporte de papel e os digitalizados poderão ser eliminados”. (BRASIL, 2018).

Assim, a grande disponibilidade de dados permite que as técnicas de visão computacional baseadas em redes neurais sejam aplicadas, já que esse tipo de algoritmo é fortemente dependente de uma grande quantidade de dados para alcançar resultados satisfatórios.

Segundo o Conselho Federal de Odontologia, hoje existem 5.472 especialistas em Radiologia Odontológica e Imaginologia (CFO, 2022). Além disso, conforme estimativas do Cfaz.net, existem cerca de 5.000 clínicas de radiologia e elas realizam em média 612 exames de traçado cefalométrico por ano. Isso representa um total de mais de 3 milhões de exames realizados por ano.

Considerando que um especialista em 2022 cobrava em média R$10,00 (dez reais) para cada exame realizado no modelo de telerradiologia, o mercado dos exames de traçado cefalométrico gera uma movimentação de cerca de R$30.000.000,00 (trinta milhões de reais) por ano, considerando apenas o Brasil. Assim como na odontologia, muitos outros exames médicos são realizados com base na análise da relação entre pontos anatômicos em radiografias. Esses exames também podem ser automatizados com as técnicas levantadas nesse trabalho.

Com isso, a motivação para a realização da presente pesquisa é o potencial de sua viabilidade, dada a obrigatoriedade legal de armazenamento de dados que possibilitem sua realização, e o fato de contribuir com uma nova ótica com as pesquisas já existentes sobre a tarefa de encontrar pontos anatômicos em imagens. Somado a esse ponto, a pesquisa traz ainda aplicabilidade, na medida em que existe um mercado financeiro extremamente atrativo para aplicação dos resultados obtidos. Além disso, existe a motivação pessoal do autor desta dissertação que se interessou pelo tema de redes neurais durante a graduação e desde então busca realizar trabalhos que permitam aplicar técnicas de aprendizado de máquina.

1.2 Definição do problema de pesquisa

A Odontologia utiliza amplamente radiografias como ferramenta tecnológica de auxílio ao diagnóstico e acompanhamento do tratamento. Assim, passou a abraçar essa nova ferramenta e a tornou indispensável para a avaliação do paciente e composição do seu prontuário médico.

A radiologia odontológica tem apresentado crescimento expressivo nos últimos anos, levando em consideração que os equipamentos de raio-x são responsáveis pela produção de radiografia. Isso, pois, segundo o estudo de Geografia Econômica da Saúde no Brasil datado de 2020, o Cadastro Nacional de Estabelecimentos de Saúde - CNES contabilizou 61.883 equipamentos de raio-x odontológicos no Brasil no ano de 2019. Esse número representa quase o dobro dos itens de assistência médica, que somam 34.128 (SALU, 2020).

O estudo também mostrou que o número de equipamentos de raio-x dentários aumentou 9,9% em apenas 2 anos, de 2017 a 2019. Para efeito de comparação, o crescimento de equipamentos de raio-x médicos foi de apenas 7,8% no mesmo período (SALU, 2020).

Esse crescimento maior na área odontológica é reflexo dos avanços tecnológicos que vem barateando o custo dos equipamentos de raio-x odontológicos, permitindo assim que dentistas tenham acesso aos equipamentos. Atualmente, cerca de um terço dos dentistas já possuem aparelhos de raio-x em seus consultórios (SALU, 2020).

No entanto, como nem todos os dentistas têm acesso aos aparelhos de raio-x, a figura da clínica de radiologia surgiu. Se refere a uma entidade que adquire os aparelhos de raio-x e atende todos os dentistas de uma região. Nesse cenário, o dentista precisa encaminhar o paciente para a realização dos exames radiográficos. A documentação ortodôntica é um conjunto de exames composto por radiografias(panorâmicas, periapicais e telerradiografia lateral), fotografias intra e extra-bucais, modelos em gesso ou impressos em resina das arcadas dentárias, análises cefalométricas e modelos computadorizados que o dentista solicita às clínicas de radiologia para ter uma visão abrangente e detalhada do caso de seu paciente. A documentação ortodôntica será utilizada no diagnóstico, planejamento, condução e avaliação do tratamento pelo dentista bem como para a sua documentação amparo legal.

O exame de traçado cefalométrico é realizado pelas clínicas de radiologia que atendem vários dentistas, de modo que esse exame precisa ser realizado diversas vezes. Essas clínicas foram as primeiras grandes consumidoras dos softwares de traçado cefalométrico, pois perceberam que utilizar o computador para calcular os fatores automaticamente diminui drasticamente o tempo para elaboração do exame e também elimina o erro humano ao efetuar os cálculos.

Com a evolução da inteligência artificial, mais especificamente da visão computacional, a radiologia odontológica está novamente à beira de um novo salto de produtividade. Nesse sentido, a definição do problema da pesquisa é que os softwares de auxílio ao diagnóstico utilizados hoje pelas clínicas de radiologia no Brasil possuem desempenho insatisfatório, levando em consideração o estado da arte das técnicas de inteligência artificial e visão computacional.

Para a definição do problema acima, foram consideradas as seguintes variáveis: (i) os softwares de cefalometria que são predominantemente empregados no mercado atual, (ii) o tempo requerido para a elaboração de um exame de traçado cefalométrico, (iii) a vivência direta do pesquisador no mercado de radiologia odontológica, e (iv) as inovações e avanços destacados em produções acadêmicas recentes na área de visão computacional.

Observa-se que as técnicas de visão computacional mais modernas são capazes de identificar pontos em imagens em tempo real. Com isso em vista, têm-se a hipótese de que a aplicação dessas técnicas para encontrar pontos cefalométricos em telerradiografias pode melhorar o desempenho dos softwares de traçado cefalométrico.

Além disso, a realidade brasileira se mostra como mais um atrativo para realização da presente pesquisa de mestrado. Isso, pelo fato de as clínicas de radiologia possuírem um campo de atuação expressivo dentro do país dado a necessidade de rateio do custo dos aparelhos de raio-x entre vários dentistas por meio de mencionadas clínicas em razão do alto valor para aquisição do equipamento de forma independente por um único dentista.

Esse modelo de negócios faz com que a clínica de radiologia acabe realizando uma grande quantidade de exames de traçado cefalométrico e assim armazene uma base de dados grande. Essa concentração de dados facilita a criação de uma base de conhecimento a ser utilizada para a presente pesquisa.

Vale ressaltar que as informações a serem utilizadas para a pesquisa necessitam ser retiradas de uma base de dados já existente uma vez que o exame de radiografia, por ser nocivo ao paciente, deve ser feito apenas quando o benefício da sua utilização suplantar o dano da sua realização, de modo que seria inviável a criação de uma base de dados com essas radiografias apenas para fins de pesquisa.

Nesse sentido os incentivos para estudos de visão computacional aplicada à radiologia odontológica são inegáveis, entretanto, o número de trabalhos na área ainda é limitado vez que um dos grandes desafios para pesquisas relacionadas a essa temática é a dificuldade na construção de uma base de dados que possa ser usada para treinamento e teste dos modelos criados para identificação dos pontos cefalométricos.

O exame de traçado cefalométrico é feito a partir da identificação de pontos cefalométricos em uma telerradiografia. Cada ponto cefalométrico é marcado em uma região definida pela literatura. Esses pontos são utilizados para realizar mensurações angulares e lineares da face, e sobre essas medidas são feitos cálculos do complexo crâniofacial do paciente permitindo que o dentista avalie seu crescimento e desenvolvimento. Há diversas maneiras de realizar o exame cefalométrico, e cada uma requer um conjunto específico de cálculos. Esse conjunto específico de cálculos, definido por cada maneira, é conhecido como análise cefalométrica.

A análise cefalométrica é então um conjunto de mensurações angulares e lineares feitas a partir dos pontos cefalométricos de uma telerradiografia. Cada medida é chamada de fator cefalométrico e uma análise cefalométrica é formada por cerca de 15 a 30 fatores. A etapa de cálculo dos fatores é determinística e já está a muito tempo resolvida por meios computacionais, restando agora a tarefa de identificação dos pontos um desafio ainda a ser resolvido por computadores de maneira satisfatória.

1.3 Objetivos

Com esse trabalho pretende-se produzir conteúdo relevante para a área de computação e radiologia odontológica, contribuindo para a melhoria na qualidade dos exames realizados ao mesmo tempo que reduz custos, proporcionando mais qualidade e eficiência no diagnóstico e plano de tratamento ortopédico/ortodôntico. Deste modo acredita-se estar contribuindo para a melhoria da saúde e qualidade de vida da população.

1.3.1 Objetivo Geral

O objetivo geral é aplicar as principais técnicas de visão computacional baseadas em redes neurais para melhoria de sistemas de auxílio ao diagnóstico, com recorte para sistema de auxílio à elaboração do exame de traçado cefalométrico.

1.3.2 Objetivos específicos

- Pesquisar o estado da arte das técnicas de visão computacional selecionando as mais aptas à tarefa de identificar pontos cefalométricos em uma telerradiografia lateral;
- Coletar uma base de conhecimento estruturada a partir de exames de traçados cefalométricos realizados por especialistas em radiologia odontológica;
- Identificar quais técnicas de visão computacional apresentam resultados satisfatórios quando utilizadas para identificar pontos cefalométricos em telerradiografias laterais;
- Avaliar como as técnicas de visão computacional estudadas se comparam ao trabalho de um especialista na elaboração do exame de traçado cefalométrico;
- Demonstrar que é possível aplicar técnicas de visão computacional através de modelos matemáticos e computacionais para auxiliar, ou até mesmo substituir, o especialista na elaboração do exame de traçado cefalométrico.

1.3.3 Resultados esperados

Através da pesquisa, pretende-se confirmar que as técnicas de visão computacional disponíveis na literatura são capazes de identificar pontos cefalométricos em telerradiografias laterais. Além disso, busca-se descobrir quais dessas técnicas são mais eficientes comparativamente para realização desta tarefa. Também pretende-se criar uma tabela comparativa mostrando como cada um dos modelos criados se comparam a especialistas humanos na realização do exame de traçado cefalométrico. Ao final, pretende-se avaliar uma estratégia que diminua o tempo de elaboração de um exame traçado cefalométrico ao mesmo tempo que produza resultados mais precisos, o que contribui para a diminuição de custos para as clínicas de radiologia e melhora do diagnóstico para o paciente.

1.4 Organização do trabalho

A dissertação está organizada em seis capítulos nos quais serão abordados os temas e assuntos relevantes ao desenvolvimento do trabalho.

No primeiro capítulo apresenta-se uma introdução ao tema para contextualizar o leitor sobre a área de radiologia odontológica, apresentar o problema que se pretende solucionar e esclarecer as motivações por trás da pesquisa. Também serão apresentados os objetivos desta.

No segundo capítulo apresenta-se toda a fundamentação teórica relacionada, a história tanto da cefalometria lateral quanto da visão computacional, para com isso levar ao conhecimento do estado da arte nessas duas áreas e apresentar como a aplicação da visão computacional pode contribuir para o exame de cefalometria lateral. Também são abordadas as principais técnicas de visão computacional com potencial para a solução do problema proposto.

No terceiro capítulo apresenta-se alguns trabalhos relacionados que foram selecionados por tratarem de temas afetos aos que estão sendo abordados nesta pesquisa.

No quarto capítulo descreve-se qual a metodologia utilizada na condução da pesquisa. Aborda-se como os dados foram coletados e estruturados, quais técnicas de visão computacional foram analisadas e explica-se porque foram selecionadas. Também é esclarecido como foi criado o traçado utilizado para referência comparativa que permite avaliar a qualidade dos pontos marcados por especialistas e pelos modelos.

No quinto capítulo apresenta-se os resultados obtidos ao atacar o problema de marcação de pontos em radiografias utilizando cada uma das técnicas de visão computacional selecionadas. Também apresenta-se a discussão sobre esses resultados.

No sexto capítulo apresenta-se as conclusões obtidas após a discussão dos resultados e o comparativo final entre os modelos e especialistas. Também sumariza-se as contribuições que a presente pesquisa proporcionou à academia, à sociedade e aos profissionais envolvidos.

2 Fundamentação Teórica

Perceber o mundo através da visão é trivial para humanos, sendo esse o principal sentido utilizado para compreensão do mundo. O ser humano é capaz de perceber variações de cores, luz e formas com facilidade. Estudiosos da visão têm tentado por décadas entender o funcionamento do mecanismo de visão e, embora já tenham conseguido criar ilusões de ótica para desvendar alguns dos princípios, o seu entendimento completo ainda não foi alcançado (MARR, 1982; WANDELL, 1995; PALMER, 1999; LIVINGSTONE, 2008; FRISBY, STONE, 2010).

Interpretação de imagens é uma habilidade fundamental para a raça humana e ao longo dos anos cada vez mais essa habilidade tem evoluído, com a criação de técnicas para obter e extrair cada vez mais informações das imagens.

Uma categoria de imagens muito antigas e de extrema importância para humanidade são as radiografias. Radiografias são imagens obtidas através da exposição de uma pessoa a radiação ionizante para a coleta de imagens de áreas internas do organismo (FREITAS, 2004). Essas radiografias são usadas desde o século XIX para a visualização de fraturas, tumores e outros males. As radiografias, em especial a telerradiografia lateral, são necessárias para auxiliar no diagnóstico, plano de tratamento e avaliação da evolução, resultado e estabilidade do tratamento ortopédico facial e/ou ortodôntico.

Pesquisas de visão computacional desenvolveram diversas técnicas matemáticas que permitem interpretar imagens e extrair delas diversas de suas características. Entretanto, apesar dos avanços no setor, o sonho de ver um computador interpretar uma imagem com o mesmo nível de profundidade que um humano parece distante. Pessoas que nunca trabalharam no campo tendem a subestimar a dificuldade dos problemas relacionados à visão computacional. Essa percepção vem da ideia ultrapassada de que problemas relacionados à lógica são mais complexos do que problemas relacionados à interpretação do ambiente (SZELISKI, 2022).

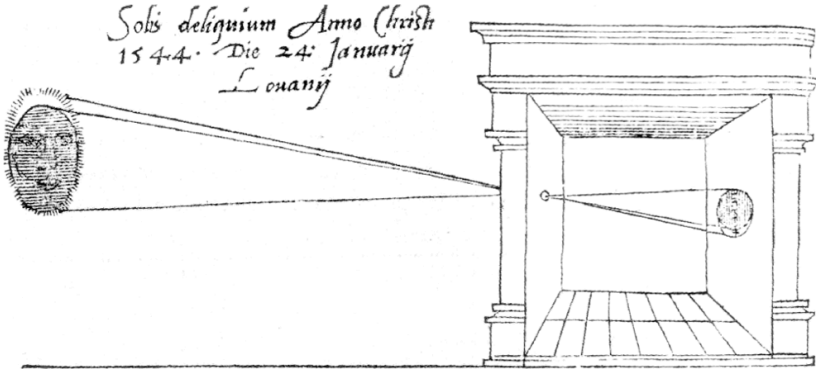
2.1 História da visão computacional

A história da visão começou há aproximadamente 543 milhões de anos atrás. Nessa época, a vida ainda era rudimentar, os organismos não tinham olhos e não se moviam muito. Quando a comida passava próximo a eles, eles simplesmente a agarravam. Mas em um curto período de tempo, aproximadamente 10 milhões de anos, o número de espécies aumentou muito, passando de alguns para centenas de milhares. Por muito tempo a causa desse aumento foi um mistério e os biólogos evolucionistas chamam esse fenômeno de big bang da evolução.

Uma das teorias mais aceitas foi proposta por Andrew Parker (MINS, 2020), que através dos estudos de fósseis descobriu nesse período que os primeiros animais desenvolveram olhos, por volta de 540 milhões de anos atrás. E foi o início da visão que permitiu esse aumento na quantidade de espécies. Com o advento da visão a vida pôde se tornar mais proativa, podendo se locomover pelo ambiente, identificar presas e evitar predadores. Depois de mais de 500 milhões de anos, a visão se tornou o sistema sensorial mais complexo na maioria dos animais, especialmente em animais inteligentes. A visão nos humanos auxilia a sobreviver, trabalhar, nos mover pelo ambiente, manipular objetos, consumir entretenimento, entre outras.

Os primeiros aparatos de visão mecânica desenvolvidos pelo homem datam de 1545, quando o físico e matemático holandês Gemma Frisius documentou um aparelho capaz de reproduzir imagens no anteparo quando a luz passava por um pequeno orifício. Esse tipo de dispositivo tem funcionamento similar aos olhos primitivos dos primeiros animais.

Figura 3 - Câmera escura de Gemma Frisius, 1545

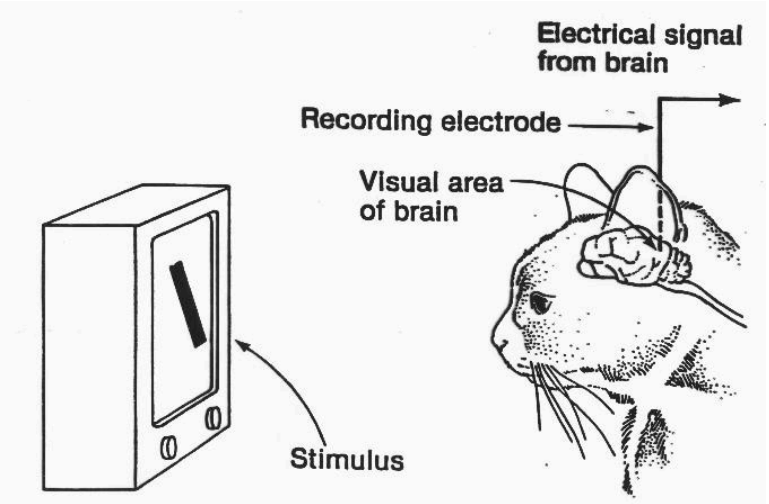


Fonte: WIKIMEDIA COMMONS, 2022

Com o passar do tempo, as câmeras evoluíram e hoje estão presentes em vários dispositivos do nosso cotidiano.

Um dos trabalhos mais influentes em visão animal foi realizado por Hubel & Wiesel (HUBEL; WIESEL, 1964). Nesse experimento eles utilizaram eletrofisiologia para tentar entender como era o processo de visão em mamíferos, eles escolheram estudar gatos. Para isso, eles colocaram eletrodos no cérebro do gato na área do córtex responsável pela visão e então observaram quais estímulos faziam os neurônios no cérebro do animal serem estimulados. O que eles descobriram foi que o processo de visão começa com reconhecimento de estruturas muito simples, se orientando através das bordas e à medida que a informação percorre o processo de visão o cérebro reconhece informações mais complexas até que ele possa reconhecer o mundo visual complexo (HUBEL; WIESEL, 1964).

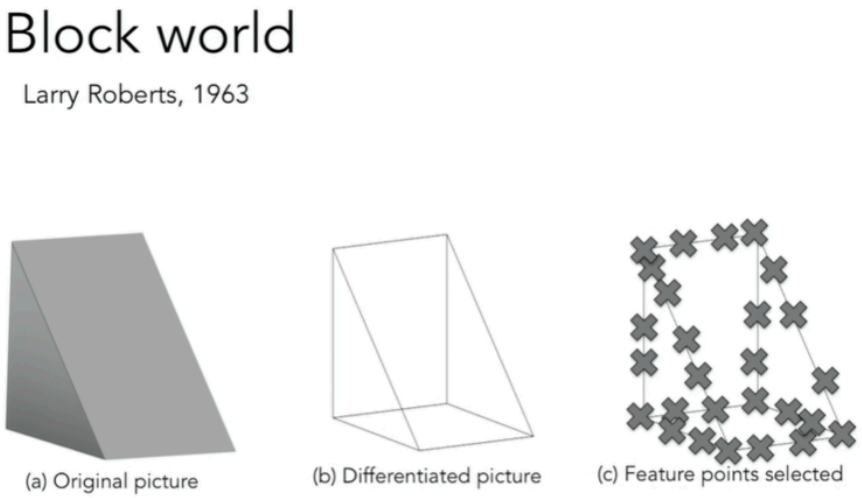
Figura 4 - Trabalho de Hubel & Wiesel



Fonte: SHU, 2013

Em 1963 Larry Roberts escreveu provavelmente o primeiro trabalho de visão computacional chamado *Block World.* Em seu trabalho, ele tentou representar o mundo visual em objetos geométricos simples e o seu objetivo era ser capaz de reconhecer o que esses objetos eram (ROBERTS, 1980).

Figura 5 - Block World de Larry Roberts



Fonte: SHU, 2013

Em 1966 o MIT realizou seu famoso projeto de verão chamado "*The Summer Vision Project*". Nesse projeto ambicioso ele pretendia utilizar a mão de obra de seus funcionários durante o verão, para realizar pesquisas, a fim de entender efetivamente partes significativas do mecanismo de visão.

Mais de 50 anos já se passaram e o campo de visão computacional passou de um projeto de verão para centenas de milhares de projetos realizados em todo o mundo. Até hoje todos os mecanismos da visão não foram desvendados, mas esse campo de estudo já se tornou em uma das mais importantes e promissoras áreas da inteligência artificial.

Outro pesquisador que merece ser mencionado é David Marr, um pesquisador do MIT que nos anos 70 escreveu um dos livros mais influentes sobre visão computacional. Em seu livro, ele descreveu o que imaginava ser a visão e como deveria ser a abordagem para construção de algoritmos de visão computacional capazes de interpretar o mundo visual. Para pegar uma imagem e a partir dela construir um modelo 3D capaz de representar o mundo, era necessário passar por diversos processos/etapas.

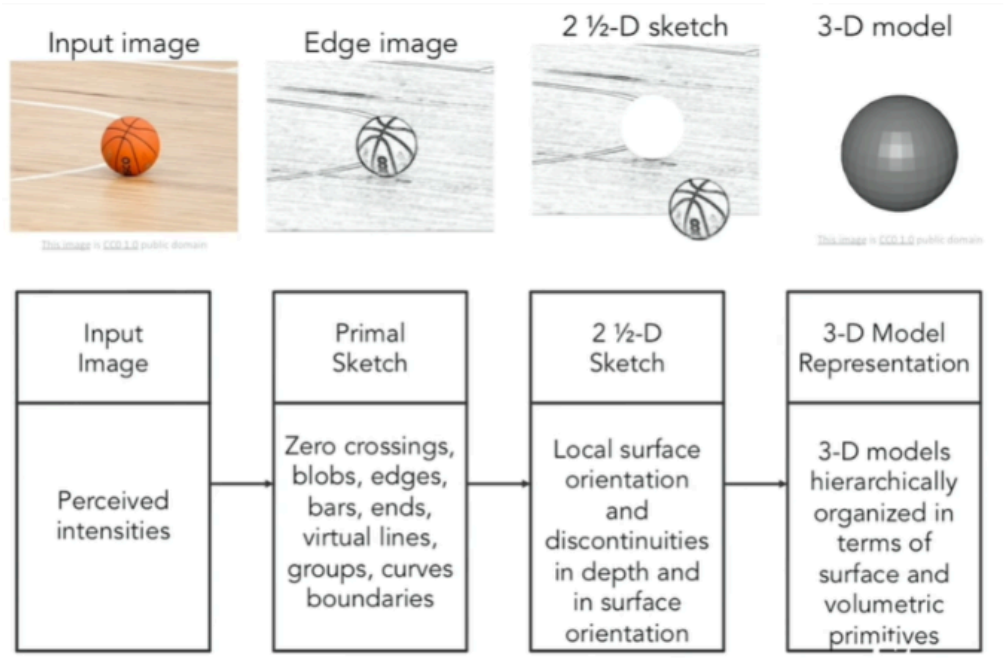
O primeiro processo foi intitulado de "*Primal Sketch*". Nessa etapa é onde ocorre o reconhecimento das bordas, curvas e limites dos objetos observados. A próxima etapa foi chamada de "*two-and-a-half-D Sketch*". Nesta, se inicia a identificação das superfícies determinadas pelas linhas e bordas, para que então seja possível identificar objetos e camadas. Já na terceira etapa, chamada de "*3-D Model Representation*", as informações

da superfície são utilizadas para estabelecer primitivas volumétricas da imagem (MARR, 1982).

Essas etapas mencionadas acima foram inspiradas por outros pesquisadores, que já imaginavam que o processo de visão se dava em fases, de modo que a construção do conhecimento sobre a imagem de forma fracionada ocorria cada vez de modo mais complexo. Essa maneira de pensar era muito intuitiva e foi o pensamento dominante na visão computacional por muitas décadas.

A imagem abaixo ilustra as etapas da visão proposta por Marr.

Figura 6 - Processo de visão proposto por Marr

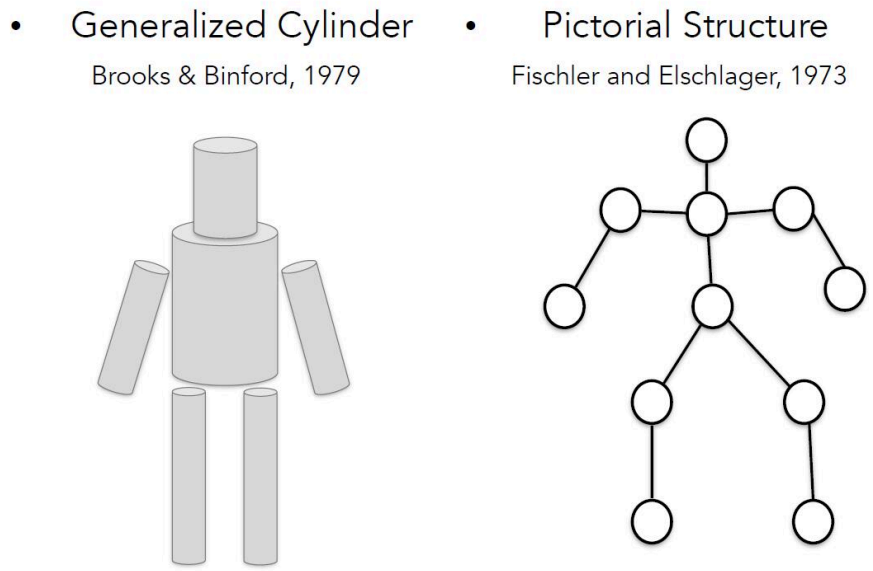


Fonte: MARR,1982

Ainda na década de 70 foram realizados trabalhos que tentavam superar a barreira de reconhecer apenas estruturas geométricas simples para reconhecer objetos complexos do mundo real. Nessa época, não havia muita disponibilidade de dados e os PC's como são conhecidos hoje ainda não existiam. Nesse período, dois grupos de pesquisadores propuseram ideias similares para reconhecimento de objetos: o *Generalized Cylinder*, proposto por Brooks e Binford em 1979, e o *Pictorial Structure,* proposto por Fischler e Elschlager em 1973. A ideia por trás dos estudos mencionados era de que qualquer objeto complexo poderia ser representado por estrutura mais

simples, como cilindros, ou então, através da distância entre pontos específicos de um objeto (PINKER, 1986). Em outras palavras, o objetivo desses estudos era simplificar objetos complexos usando uma composição de objetos mais simples.

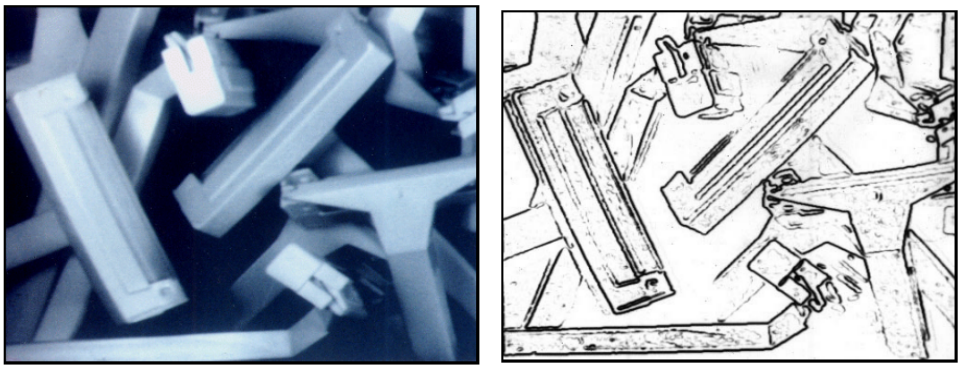
Figura 7 - Representação de objetos complexos



Fonte: MINS, 2020

Nos anos 80 o pesquisador David Lowe utilizou técnicas de visão computacional similares para identificar objetos complexos. Em sua pesquisa, ele tentou utilizar operações de convolução sobre a imagem para identificar as bordas relevantes dessa e, a partir das obras produzidas, identificar os objetos observados (LOWE, 1987).

Figura 8 - Convolução para identificação de bordas



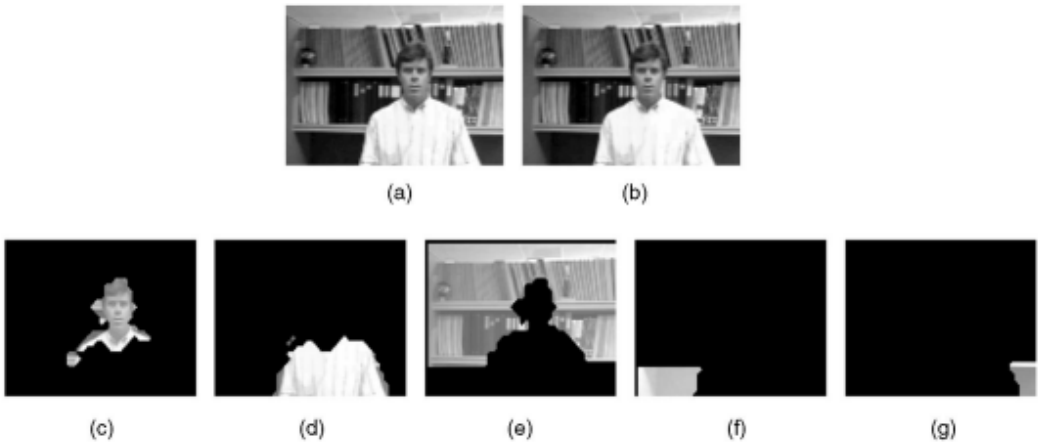
Fonte: LOWE, 1987

Até os anos 80 houve muito esforço na tarefa de identificação de objetos. Entretanto, essa atividade ainda se provava muito complexa, sendo difícil a obtenção de resultados satisfatórios. Todos os trabalhos eram realizados sobre uma base de dados pequena e não apresentavam resultados que poderiam ser usados em larga escala no mundo real.

Então, percebendo que era difícil completar a tarefa de reconhecimento de objetos, pesquisadores resolveram tentar fazer a segmentação destes. Essa tarefa consistia em utilizar uma imagem e separar os pixels em áreas de relacionamento semântico, ou seja, segmentar a imagem. Vale dizer que a segmentação de imagens é um campo de visão computacional que possui diversos trabalhos relacionados.

Um dos primeiros trabalhos de segmentação de imagem foi o Normalized Cut (SHI; MALIK, 2000). No desenvolvimento de sua pesquisa, ele utilizou um algoritmo de particionamento de grafo para segmentar a imagem. A figura abaixo mostra um exemplo da segmentação realizada.

Figura 9 - Exemplo de segmentação de imagem



Fonte: SHI; MALIK, 2000

Outro problema que ganhou notoriedade no campo de visão computacional foi a tarefa e desconhecimento de faces. Faces são um dos objetos mais importantes para os seres humanos, e o reconhecimento destas possui inúmeras aplicações na sociedade como é conhecida.

Um dos primeiros trabalhos que se propôs a realizar o reconhecimento facial foi realizado por Viola e Jones. Em seus estudos, eles propuseram um algoritmo capaz de reconhecer a face de humanos através de um algoritmo computacionalmente eficiente. Eles foram capazes de reconhecer faces em imagens de 384x288 pixels utilizando um pentium de 700 MHz (VIOLA; JONES, 2001).

Poucos anos depois da publicação do estudo tratado acima, em 2006, a empresa FujiFilm lançou a primeira câmera fotográfica capaz de fazer o reconhecimento facial em tempo real. Esta foi uma transição realmente rápida entre pesquisa científica e aplicação do mundo real. A figura abaixo ilustra a utilização do algoritmo de Viola e Jones em uma imagem com uma face.

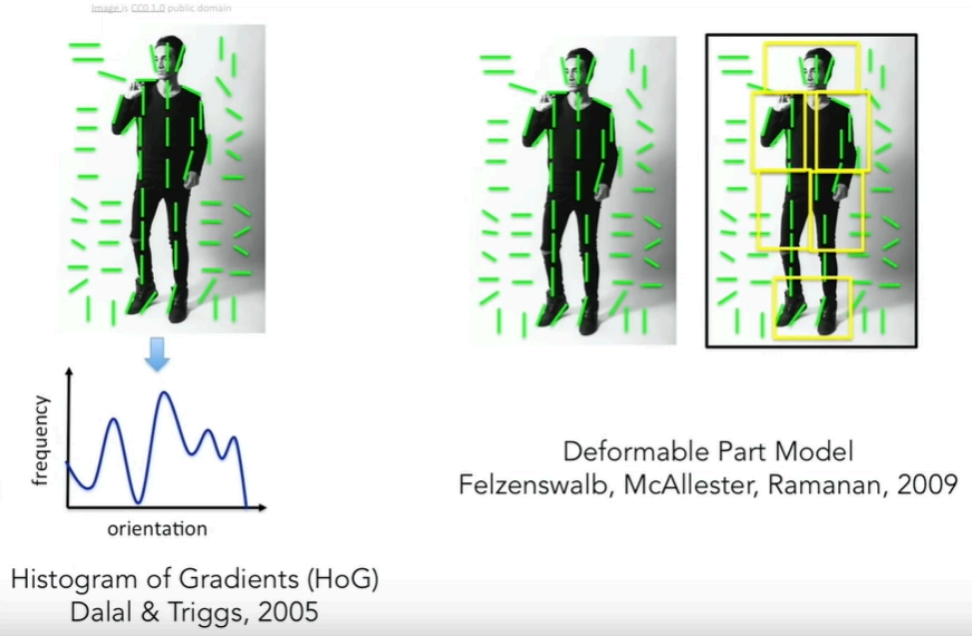
Figura 10 - Algoritmo Viola-Jones para reconhecimento facial

Fonte: VIOLA; JONES, 2001

Com o passar dos anos houve a evolução nas técnicas de reconhecimento de objetos. Nos anos 2000, começaram a surgir trabalhos que faziam o reconhecimento de imagem através da extração de características das imagens. Um dos principais algoritmos foi o SIFT (*Scale-Invariant Feature Detector*), desenvolvido por David Lowe em 1999. Esse algoritmo consiste em encontrar pontos fundamentais nas imagens chamadas de *keypoints* e utilizar esses pontos como uma assinatura da imagem. O algoritmo tinha bons resultados identificando objetos que foram observados em diferentes ângulos e posições, pois esses pontos fundamentais permanecem invariantes entre as imagens.

Nesse mesmo período surgiram outros trabalhos que tentaram reconhecer corpos humanos utilizando o histograma da orientação dos gradientes para identificar objetos na imagem (DALAS; TRIGGS, 2005; FELZENSZWALB; GIRSHICK; MCALLESTER, 2010).

Figura 11 - Histograma de gradientes



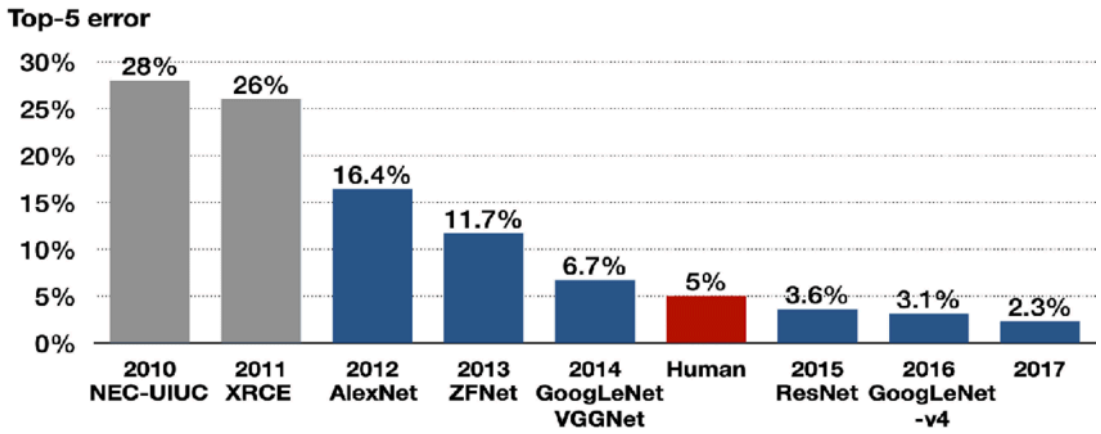
Fonte: MINS, 2020

Foi também nos anos 2000 que a existência de bancos de dados de imagens para *benchmark* teve início. O banco de dados de imagens permitiu a medição do progresso do reconhecimento de imagem. Um dos primeiros banco de imagens criados para fazer o *benchmark* para classificação de imagens foi o *PASCAL Visual Object Challenge*. O banco de imagens contava com 20 categorias, com milhares de imagens para cada categoria.

Outro banco de imagens relevante para a evolução da visão computacional foi o ImageNet. Esse banco de imagens conta com mais de 14 milhões de imagens organizadas em 22 mil categorias. O grupo responsável pelo ImageNet criou em 2009 um desafio internacional para reconhecimento de imagens. Neste desafio, foram selecionadas 1,4 milhões de imagens e mil classes para que pesquisadores pudessem testar e comparar os algoritmos de reconhecimento de imagem. O desafio consistia em identificar com sucesso 5 categorias incluídas na imagem para cada uma das imagens

selecionadas. O gráfico abaixo mostra o desempenho dos algoritmos de classificação de imagem.

Figura 12 - Resultado do ImageNet Challenge



Fonte: KANG, DUONG, PARK, 2020

No gráfico verifica-se a melhoria drástica que ocorreu em 2012 com o surgimento da AlexNet. Considerada a primeira rede neural convolucional utilizada para a tarefa de classificação de imagens, essa arquitetura de rede alcançou resultados superiores aos algoritmos tradicionais de visão computacional. A partir de 2012, uma nova fase da visão computacional teve início, na qual redes neurais convolucionais dominaram como a melhor alternativa para tarefas de visão computacional.

Atualmente, os melhores modelos para tarefa de classificação de imagens são as variações da ResNet, apresentadas inicialmente em 2015. ResNet se refere a um modelo que usa Redes Neurais Convolucionais (CNNs) para classificar imagens. O ranking atual de modelos de classificação de imagem pode ser visto no site https://paperswithcode.com/sota/image-classification-on-imagenet.

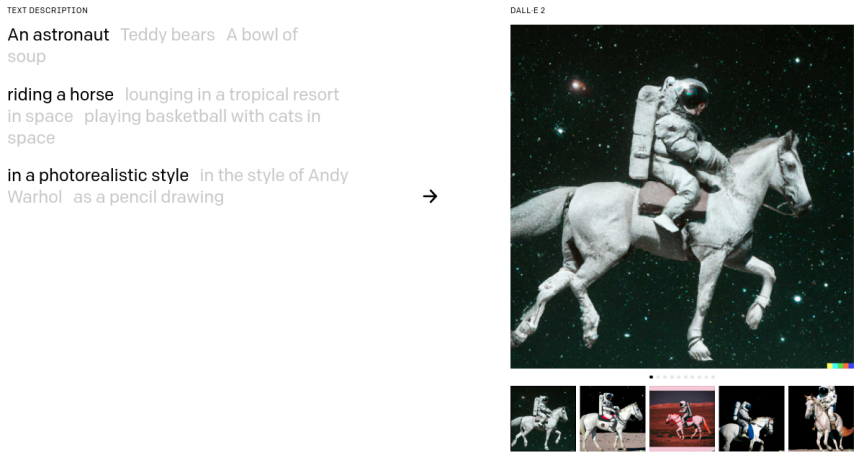
Em outubro de 2020, Alexey Dosovitskiy apresentou pela primeira vez o conceito de Vision Transformers (ViT) para tarefas de classificação de imagem. Transformers foram introduzidos em 2017 e até então só possuíam aplicação para processamento de linguagem natural (VASWANI, 2017). Em seu seu artigo "*An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*", Alexey mostrou que era possível aplicar a técnica em tarefas de classificação de imagens e obteve resultados que superam os resultado obtidos com a ResNet.

Para que os resultados com ViT sejam superiores aos da ResNet, a quantidade de imagens para treinamento precisa ser muito grande. Em sua pesquisa, Alexey comparou os resultados entre a ResNet e a ViT. Ele observou que, para até 100 milhões de imagens, a ResNet teve resultado superior que a ViT; para 100 milhões de imagens, os resultados foram similares; e para 300 milhões de imagens, os resultados com ViT foram superiores aos da ResNet (DOSOVITSKIY, 2020). Vale dizer que o banco de imagens utilizado na pesquisa foi o JFT da Google, não disponível publicamente e que conta com 300 milhões de imagens. Isso porque, o banco de imagens da ImageNet possui apenas 14 milhões de imagens.

Em 2022, o modelo que obteve o melhor resultado no desafio de classificação de imagens no banco de dados ImageNet foi o *Contrastive Captioners are Image-Text Foundation Models* (CoCa). Esse modelo utiliza *vision transformers* para realizar a tarefa de classificação de imagens.

Outra área da visão computacional que merece ser mencionada é a rede Generative Adversarial Networks (GAN). Ela foi proposta em 2014 por Ian Goodfellow e, diferente das técnicas mencionadas anteriormente, não possui o objetivo de reconhecer imagens, mas sim de gerá-las (GOODFELLOW, 2014). Uma das suas aplicações mais conhecidas está na geração de rostos humanos. Recentemente, esse tipo de rede ganhou popularidade ao permitir que imagens sejam geradas a partir de entradas de texto ou a partir de outras imagens. O site https://openai.com/dall-e-2/ é um dos modelos mais impressionantes e permite que qualquer pessoa possa usá-lo gratuitamente.

Figura 13 - Geração de imagem utilizando o Dall-E 2



Fonte: OPENAI, 2022

2.2 História da cefalometria lateral

A primeira radiografia foi realizada em 1859 pelo professor Wilhelm Conrad Röntgen. Enquanto fazia experimentos com raios catódicos, ele percebeu que esses raios eram capazes de gerar imagens das estruturas internas do organismo. Para isso, era necessário expor o organismo aos raios contra um filme fotográfico por alguns minutos. Esses raios foram chamados de raios X (LANGLAND; LANGLAIS, 1997).

Figura 14 - 1ª Radiografia da história – mão de Anna Bertha esposa de Röntgen

Fonte: MUSEUM, 2022

A descoberta das radiografias possibilitou um grande avanço para a odontologia ao passo em que permitiu melhor compreensão das estruturas internas do corpo e uma grande evolução dos diagnósticos. Desse modo, as radiografias contribuíram para o surgimento e ampliação de novas áreas na odontologia (MARTINS, 2005).

A primeira pessoa a utilizar radiografias na odontologia foi o Dr. Edmund Kells em 1899. Ele escreveu um trabalho publicado no periódico “Dental Cosmos”, em agosto de

1899. Em seu trabalho, a importância de se tomar uma radiografia com ângulos corretos e posicionadores para o filme radiográfico é descrita. Isso acontece porque, dependendo da angulação do filme durante a aquisição, podem haver distorções na imagem (MARTINS, 2005).

Após realizar incontáveis radiografias sem a devida proteção, Kells descobriu na prática quais são os efeitos deletérios da radiação. Ele perdeu inicialmente os dedos da mão esquerda e, em seguida, perdeu a mão. Mais tarde perdeu também o braço esquerdo que teve de ser amputado. Apesar disso, continuou trabalhando com odontologia e desenhos de aparelhos que o permitiram trabalhar apenas com a mão direita. Entretanto, com o passar do tempo, sua mão direita também foi afetada pela radiação. Sua luta contra os efeitos da radiação durou 20 anos, período esse em que se submeteu a várias cirurgias até o dia que cometeu autoextermínio (MARTINS, 2005).

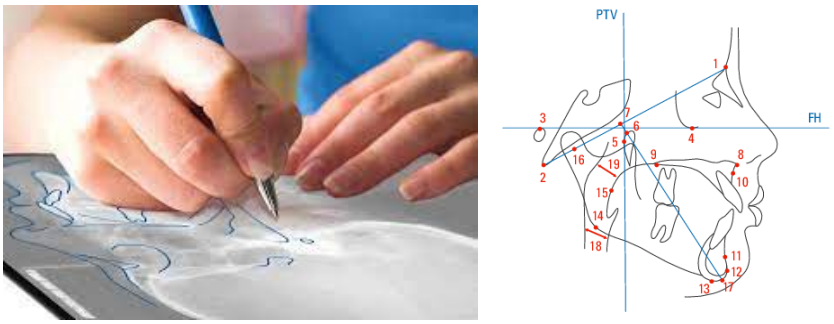
Vale a pena ressaltar que Broadbent, desenvolveu o cefalostato em 1931, dispositivo que permitiu a padronização da posição da cabeça, e consequentemente permitiu a obtenção de telerradiografias laterais da cabeça padronizadas que puderam ser sobrepostas. A partir desse momento foi possível utilizar estas imagens para realização de estudos de crescimento craniofacial sobrepondo as radiografias e avaliando os incrementos ósseos e direção de crescimento. Seu trabalho foi publicado no artigo "*A new X-ray technique and its application to orthodontics*" no Congresso da Chicago Dental Society, em 4 de fevereiro de 1931.

A radiografia lateral da cabeça, ou telerradiografia lateral, é utilizada para a realização do cefalograma. Esse, por sua vez, consiste do desenho anatômico juntamente com mensurações lineares e angulares dos pontos cefalométricos. Por pontos cefalométricos, entende-se regiões precisamente especificadas por autores em uma radiografia lateral da face (telerradiografia) na qual, além dos pontos, também são analisados planos, que são obtidos com base na marcação dos pontos (VILELLA, 2009).

Inicialmente, o traçado cefalométrico foi realizado manualmente através da marcação de pontos sobre uma folha de papel de acetato, brilhante do lado que entra em contato com a radiografia e do outro lado a textura permite o desenho com facilidade. Os pontos eram identificados e marcados manualmente, depois as linhas e planos eram traçados, e posteriormente os ângulos e medidas lineares eram mensurados com o auxílio de réguas, transferidores de ângulo e esquadros. As medidas obtidas eram

comparadas aos padrões (referências) para auxiliar no diagnóstico e planejamento do tratamento ortopédico facial e/ou ortodôntico.

Figura 15 - Desenho anatômico, pontos, linhas e planos demarcados manualmente



Fonte: VILELLA, 2009

Ao resultado das mensurações desses pontos e planos dá-se o nome de fatores cefalométricos. Esses fatores, juntamente com as respectivas normas e desvios padrões, são apresentados em uma tabela de resultados. A Figura 16 mostra parte de uma tabela de fatores proposta por Ricketts (VILELLA, 2009).

Figura 16 - Conjunto de fatores cefalométricos de Ricketts

#	Descrição	Valor	Padrão
Campo I - Problemas Dentários			
1	Relação Molar	-1.36 mm	-3.0 ± 3.0
2	Relação Canina	0.168 mm	-2.0 ± 3.0
3	Trespasse Horizontal do Incisivo (Overjet)	2.736 mm	2.5 ± 2.5
4	Trespasse Vertical do Incisivo (Overbite)	2.587 mm	2.5 ± 2.0
5	Extrusão do Incisivo Inferior	1.68 mm	1.25 ± 2.0
6	Ângulo Interincisivo	135.332 gr	130.0 ± 6.0

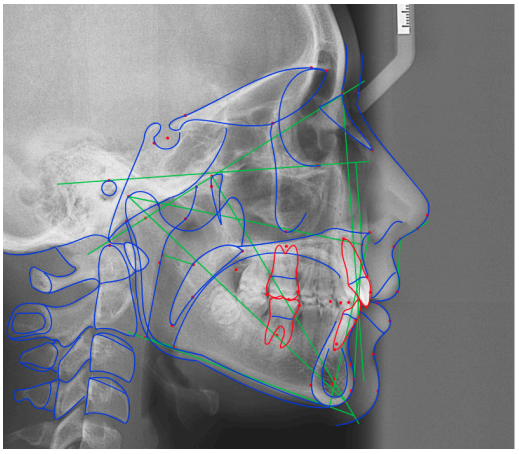
Fonte: CFAZ.NET, 2020

Ao conjunto de pontos, linhas, planos e fatores observados dá-se o nome de análise cefalométrica.

Os avanços tecnológicos no campo da informática permitiram a realização de radiografias digitais, que agora passaram a ser geradas pela captura dos raios-X diretamente por sensores, o que diminuiu a dose de radiação necessária e o tempo de exposição necessário para a geração das radiografias. Além disso, filtros puderam ser aplicados melhorando a qualidade da radiografia (CHEN, 2014).

A partir da década de 60, computadores começaram a ser utilizados para marcar os pontos cefalométricos de forma digital, permitindo assim que as medidas e cálculos pudessem ser computados instantaneamente (PAIVA, 2008). A utilização de computadores contribuiu para a popularização da técnica cefalométrica, que hoje é largamente utilizada no mercado de odontologia.

Figura 17 - Desenho anatômico, pontos, linhas e planos cefalométricos



Fonte: CFAZ.NET, 2020

Nos últimos anos, os avanços no campo da visão computacional estão permitindo um novo passo para melhorar e baratear o exame de traçado cefalométrico, permitindo que os pontos anatômicos possam ser realizados de forma automática. Vale dizer que já existem trabalhos relacionados ao dito e que os resultados são promissores (SONG; QIAO; IWAMOTO, 2020).

2.3 Análise cefalométrica no padrão USP

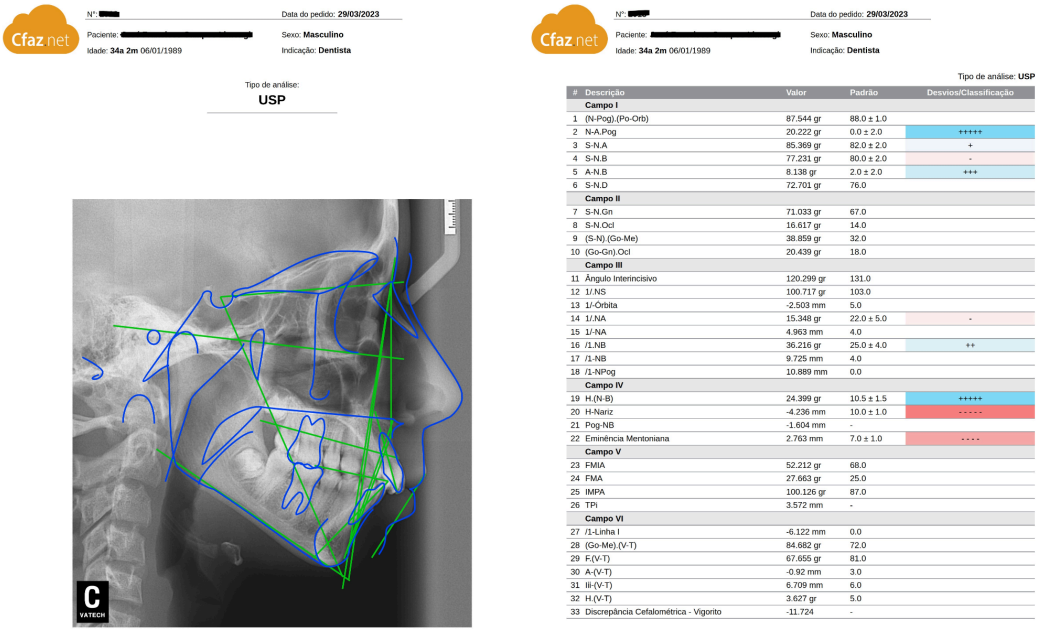
A análise cefalométrica no padrão USP foi criada pelo curso de Pós-graduação de Ortodontia da Faculdade de Odontologia da Universidade de São Paulo - FOUSP em 1968. Essa análise consiste em uma seleção de fatores cefalométricos originários de diversas análises cefalométricas em norma lateral. Para criação da análise no padrão USP foram utilizadas as análises de Steiner, Downs, Tweed, Interlandi e Holdaway (INTERLANDI, 1968). Desde sua concepção o padrão criado ganhou popularidade e é amplamente utilizado por profissionais de odontologia.

A análise no padrão USP foi a mais utilizada no Brasil em exames de traçado cefalométrico, conforme avaliação de dados da plataforma Cfaz.net entre 2015 e 2020. Neste intervalo, constatou-se que aproximadamente 35% dos exames empregaram esta análise, destacando sua predominância e relevância. Assim, sua aplicação neste estudo é adequadamente justificada.

A análise consiste na avaliação desses diversos fatores cefalométricos que são comparados com valores de referência estabelecidos. Observando esses fatores podemos obter informações importantes sobre a estrutura do craniofacial do paciente. Com a análise no padrão USP podemos avaliar se o perfil da face é reto, côncavo ou convexo, se existe protrusão ou retrusão da maxila e mandíbula, podemos avaliar a relação esquelética entre maxila e mandíbula (Classe I,II ou III esquelética) e também podemos avaliar a direção predominante de crescimento da face, especialmente da mandíbula em relação a uma linha (linha S-N) ou a um plano de referência (plano de Frankfurt - PoOr). Essas informações são utilizadas pelo ortodontista para elaborar um diagnóstico e o melhor plano de tratamento para o paciente.

A figura 18 mostra uma análise no padrão de USP, nela podemos observar o desenho das estruturas cefalométricas sobre a telerradiografia e a tabela com a avaliação dos fatores cefalométricos.

Figura 18 - Análise no padrão USP



Cfaz net

Nº: [redacted] Data do pedido: **29/03/2023**

Paciente: [redacted] Sexo: **Masculino**

Idade: **34a 2m** 06/01/1989 Indicação: **Dentista**

Tipo de análise: **USP**

Cfaz net

Nº: [redacted] Data do pedido: **29/03/2023**

Paciente: [redacted] Sexo: **Masculino**

Idade: **34a 2m** 06/01/1989 Indicação: **Dentista**

Tipo de análise: **USP**

#	Descrição	Valor	Padrão	Desvios/Classificação
	Campo I			
1	(N-Pog).(Po-Orb)	87.544 gr	88.0 ± 1.0	
2	N-A.Pog	20.222 gr	0.0 ± 2.0	+++++
3	S-N.A	85.369 gr	82.0 ± 2.0	+
4	S-N.B	77.231 gr	80.0 ± 2.0	-
5	A-N.B	8.138 gr	2.0 ± 2.0	+++
6	S-N.D	72.701 gr	76.0	
	Campo II			
7	S-N.Gn	71.033 gr	67.0	
8	S-N.Ocl	16.617 gr	14.0	
9	(S-N).(Go-Me)	38.859 gr	32.0	
10	(Go-Gn).Ocl	20.439 gr	18.0	
	Campo III			
11	Ângulo Interincisivo	120.299 gr	131.0	
12	1/.NS	100.717 gr	103.0	
13	1/-Órbita	-2.503 mm	5.0	
14	1/.NA	15.348 gr	22.0 ± 5.0	-
15	1/-NA	4.963 mm	4.0	
16	/1.NB	36.216 gr	25.0 ± 4.0	++
17	/1-NB	9.725 mm	4.0	
18	/1-NPog	10.889 mm	0.0	
	Campo IV			
19	H.(N-B)	24.399 gr	10.5 ± 1.5	+++++
20	H-Nariz	-4.236 mm	10.0 ± 1.0	- - - - -
21	Pog-NB	-1.604 mm	-	
22	Eminência Mentoniana	2.763 mm	7.0 ± 1.0	- - - -
	Campo V			
23	FMIA	52.212 gr	68.0	
24	FMA	27.663 gr	25.0	
25	IMPA	100.126 gr	87.0	
26	TPi	3.572 mm	-	
	Campo VI			
27	/1-Linha I	-6.122 mm	0.0	
28	(Go-Me).(V-T)	84.682 gr	72.0	
29	F.(V-T)	67.655 gr	81.0	
30	A-(V-T)	-0.92 mm	3.0	
31	Ii-(V-T)	6.709 mm	6.0	
32	H.(V-T)	3.627 gr	5.0	
33	Discrepância Cefalométrica - Vigorito	-11.724	-	

Fonte: CFAZ.NET, 2023

2.4 Redes neurais

Redes neurais são sistemas computacionais inspirados no funcionamento do cérebro humano, capazes de aprender a partir de exemplos e tomar decisões de forma autônoma. Redes neurais não são nenhuma novidade, os primeiros trabalhos relacionados datam de 1957, quando Frank Rosenblatt fez os primeiros experimentos com um perceptron.

Nos últimos anos, as redes neurais se tornaram uma das técnicas mais promissoras em áreas como visão computacional, reconhecimento de fala, processamento de linguagem natural, entre outras. Em particular, na visão computacional, as redes neurais têm sido amplamente utilizadas para tarefas como reconhecimento de objetos, detecção de faces, classificação de imagens e detecção de pontos chave em imagens, como é o caso da detecção de pontos cefalométricos em telerradiografias.

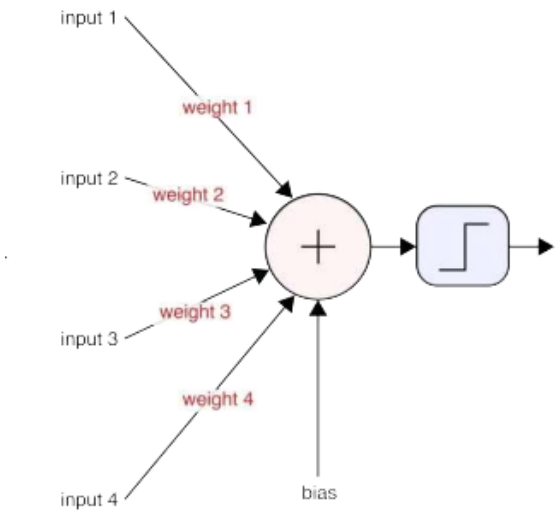
Neste capítulo serão apresentados os principais conceitos relacionados às essas redes neurais e como elas podem ser aplicadas a problemas de visão computacional.

2.4.1 Perceptron de Multicamadas (MLP)

A estrutura central de processamento de uma rede neural é o perceptron. Essa estrutura, que foi concebida para tentar simular um neurônio, é a peça central na construção de todas as redes neurais que conhecemos hoje.

Na Figura 19, que representa o perceptron, podemos ver uma lista de parâmetros de entrada sendo multiplicados por pesos e somados a um bias. Esse valor então é fornecido a uma função de ativação simples que retorna 1 se o resultado dessa operação for valor maior que zero ou 0 caso contrário.

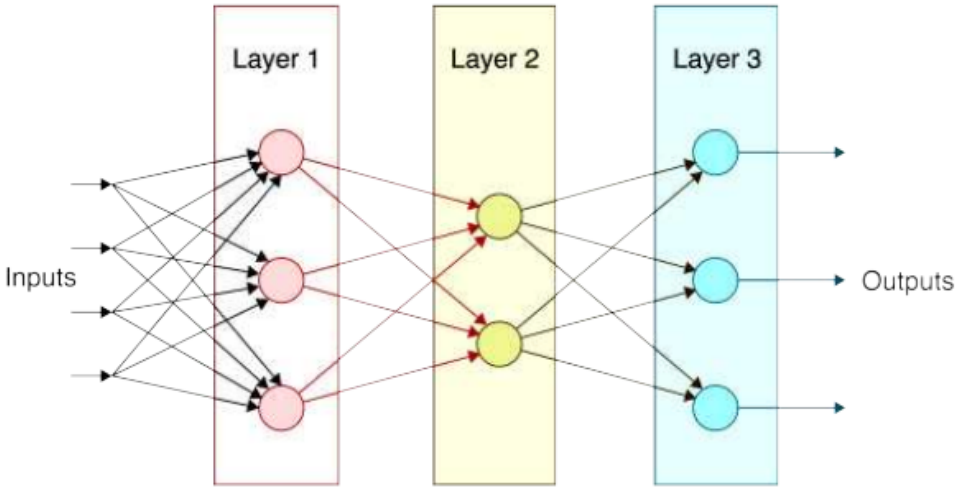
Figura 19 - Representação gráfica de um perceptron



Fonte: SZELISKI, 2022

Os perceptrons podem ser usados em conjunto, ao se utilizar mais de um perceptron conectado aos dados de entrada teremos uma camada. Adicionando mais camadas que utilizam como dados de entrada a saída da camada anterior obtemos uma rede neural de multicamadas também conhecida como *deep neural network*. A saída da última camada da rede neural representa o resultado que queremos extrair dos dados de entrada. A figura 20 ilustra a arquitetura de uma rede neural de multicamadas.

Figura 20 - Rede Neural de Multicamadas



Fonte: SZELISKI, 2022

O resultado da rede neural depende diretamente dos pesos utilizados na operação, são eles que juntamente à arquitetura da rede representam a inteligência dessa rede neural.

Os valores ideais dos pesos são descobertos através do processo de treinamento da rede neural. Nesse processo são utilizados dados de entrada e saída de uma base de dados que são exaustivamente processados em um processo de treinamento. Durante o treinamento os valores dos pesos são inicialmente aleatórios e são alterados a cada iteração do treinamento a fim de diminuir o erro da rede neural, ou seja encontrar os valores dos pesos que melhor processam os dados de entrada para os valores de saída da base de dados.

O erro do processamento dos dados de entrada pode ser calculado através de uma função de erro. Existem diversas funções de erro, por exemplo, a soma dos quadrados das diferenças dos valores de saída computados pela rede e o valor de saída existente na base de dados.

Para saber a melhor forma de ajustar o valor dos pesos da rede neural deve-se calcular o gradiente desta função de erro através da derivada parcial da função de erro em relação aos pesos. Assim, a cada iteração do treinamento ajustam-se os valores dos pesos na direção oposta ao gradiente da função de erro. Para isso é utilizado o algoritmo de *backpropagation*.

O algoritmo de *backpropagation* é o algoritmo utilizado para calcular o gradiente em cada camada da rede, a partir da camada de saída até a camada de entrada. Esse algoritmo utiliza a regra da cadeia para calcular as derivadas parciais em cada camada, de forma eficiente. À magnitude da alteração nos pesos a cada etapa do treinamento damos o nome de taxa de aprendizado (SZELISKI, 2022).

Uma forma de utilizar redes neurais de multicamadas para problemas de visão computacional é criando um vetor que inclua todos os pixels da imagem e alimentar a primeira camada da rede com esses valores como entrada. Essa abordagem perde as característica espaciais da imagem, pois todos os pixels foram removidos de posição e adicionados em um array. Ainda assim, esse método consegue assimilar informações das imagens e normalmente é suficiente para problemas de classificação.

2.4.2 Rede neural convolucional (CNN)

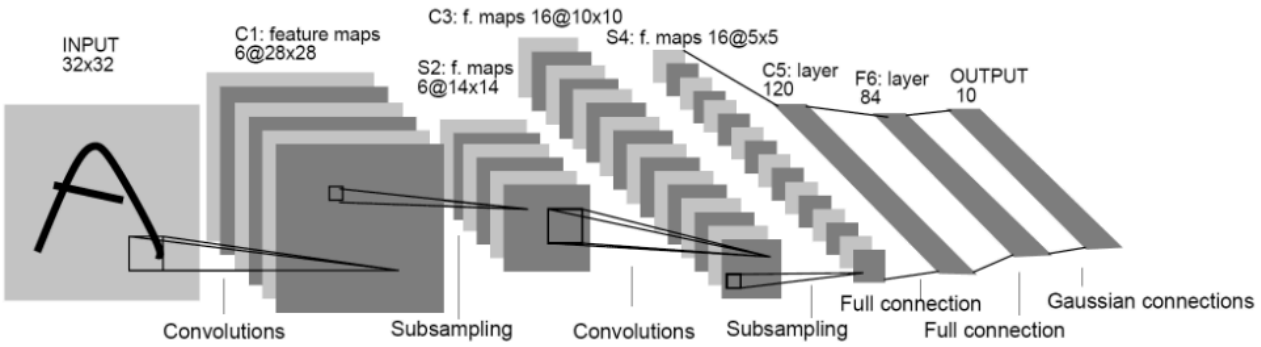
As CNNs são uma arquitetura de rede neural normalmente usada para tarefas de visão computacional, como classificação de objetos, detecção de objetos e segmentação de objetos. Assim como a rede neural de multicamadas, ela deve ser treinada em uma base de dados para calibrar seus pesos.

Para processar uma imagem a rede neural convolucional utiliza uma matriz de pesos chamada kernel, normalmente de tamanho 3x3, que deve ser multiplicada por todos os pixels da imagem, deste modo, fazendo uma varredura de toda a imagem e gerando um mapa de características para essa imagem. O mapa de características é uma versão mais compacta e significativa das características importantes dessa imagem. Esse processo de multiplicação da imagem pelo kernel é conhecido como convolução e deve pode ser repetido em múltiplas camadas para melhorar o desempenho da rede.

Além das camadas de convolução, a CNN também possui camadas de *pooling* que realizam uma amostragem da imagem diminuindo a sua resolução e consequentemente aumentando o desempenho da rede. As últimas camadas da CNN são compostas por neurônios completamente conectados para realizar a classificação final da rede.

A Figura 21 mostra uma rede neural convolucional concebida para identificação de dígitos. Essa rede usa múltiplas camadas de convolução, realiza a redução de resolução e também camadas completamente conectadas para classificar uma imagem de entrada de tamanho 32x32 em um dos 10 dígitos na saída.

Figura 21 - Rede Neural Convolucional



Fonte: SZELISKI, 2022

Redes neurais convolucionais obtiveram os melhores resultados para tarefas relacionadas à visão computacional em 2012 com a introdução da *AlexNet*, uma rede neural de 9 camadas utilizada para classificação de imagens. Desde então, novas redes de camadas ainda mais profundas foram desenvolvidas com resultados ainda melhores, como a VGG e a GoogleNet.

As redes neurais convolucionais foram as principais ferramentas para tarefas de visão computacional até 2020, quando os transformers começaram a ser utilizados também para tarefas de visão computacional com resultados similares aos resultados das CNNs.

2.4.3 Rede Neural Recorrente (RNN)

Redes neurais recorrentes podem ser usadas quando queremos que o tamanho da entrada da rede seja diferente do tamanho da saída da rede. Esse recurso é particularmente interessante quando queremos processar linguagem, por exemplo, para tradução de texto. Nesse caso, a tradução de uma sentença de entrada em inglês tem tamanho diferente da sequência de saída em português. Também para um modelo de tradução é importante que o tamanho da sentença de entrada também precise ser variável.

Em geral, uma RNN tem um vetor de entrada X que é usado para alimentar a RNN. A RNN tem estados ocultos internos que são atualizados a cada leitura de uma entrada do vetor X. A cada nova leitura de uma entrada do vetor X um estado também vai ser passado para a RNN e um novo estado vai ser gerado. A RNN tem um vetor de saída Y. Assim temos um padrão de funcionamento no qual uma entrada é lida, o estado é atualizado e uma saída é produzida a cada etapa. A cada etapa da computação é usada a mesma função e os mesmos pesos para o processamento (SZELISKI, 2022).

As fórmulas abaixo exemplificam o funcionamento de uma RNN.

A fórmula (1) mostra a estrutura matemática da rede. Sendo: h é o estado; x é o vetor de entrada; f é a função aplicada e w a matriz de pesos.

$$h_t = fw(h_{t-1}, x_t) \tag{1}$$

Com essa estrutura podemos pensar em uma matriz de pesos W_{xh}que pode ser multiplicada pela entrada x_t e uma matriz de peso W_{hh} que pode ser multiplicada pelo estado h_{t-1}. Essas matrizes são então combinadas e então é aplicada uma tangente hiperbólica para adicionar não-linearidade ao modelo. Conforme a fórmula (2):

$$tanh(W_{hh}h_t + W_{xh}x_t) \quad (2)$$

Além disso, se você deseja ter uma previsão a cada etapa, você também pode ter outra matriz W_{ht} que utilize os pesos multiplicado pelo valor h_t para obter um saida y_t. Conforme fórmula (3):

$$y_t = W_{ht}h_t \quad (3)$$

Uma variação das redes neurais RNN que vale a pena ser citada são as Long Short Term Memory (LSTM). Elas são um tipo de rede recorrente muito usada para processamento de linguagem natural (SZELISKI, 2022).

Uma aplicação das RNN em imagens é para criação de legendas para essas imagens. Para isso é necessário levar em consideração as informações da imagem durante o processamento da rede. Isso pode ser alcançado adicionando uma nova matriz de pesos W_{ih} para adicionar informações da imagem e utilizá-la para calcular o próximo estado oculto durante as interações. Realizando essa alteração a fórmula de cálculo do próximo estado oculto seria:

$$h_t = tanh(W_{hh}h_t + W_{xh}x_t + W_{ih}v) \quad (4)$$

Apesar de muito úteis para processamento de linguagem natural e existirem algumas aplicações em imagens, as RNN não são as redes mais indicadas para processamento de imagem, portanto o estudo do seu funcionamento não foi aprofundado

neste trabalho, visto que a sua utilização não se mostrou promissora para resolver o problema de identificação de pontos anatômicos em radiografias.

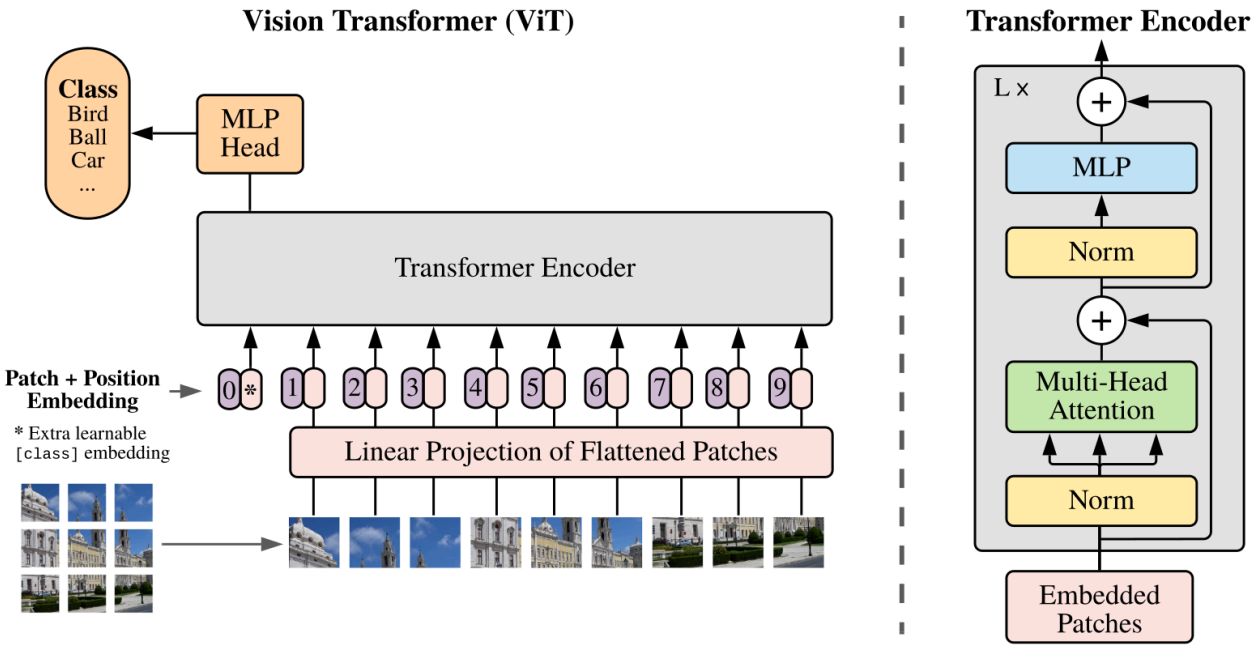
2.4.4 Vision Transformers (ViT)

A introdução do mecanismo de self-attention foi realizada por Ashish Vaswani em seu artigo "Attention Is All You Need" (VASWANI, 2017). Esse mecanismo é aplicado ao processamento de linguagem natural (PLN) e permite que as redes neurais identifiquem quais palavras são mais relevantes em uma frase, melhorando significativamente sua capacidade de compreensão e interpretação. A estrutura de redes neurais proposta por Vaswani, ficou conhecida como transformers e tem sido amplamente utilizada em modelos de PLN como o Chat-GPT da OpenAI e o BERT do Google. Os transformers permitiram avanços significativos na geração de linguagem natural e em outras tarefas de processamento de texto.

Visions Transformers (ViT) é uma arquitetura muito semelhante ao modelo de transformers para tarefas de visão computacional, como classificação de imagens e detecção de objetos. Com pequenas alterações para processar imagens em vez de textos. A abordagem ViT tem mostrado resultados promissores em relação às arquiteturas convencionais de CNNs (Convolutional Neural Networks), que foram o estado da arte em tarefas de visão computacional.

A arquitetura ViT foi proposta pela primeira vez em um artigo intitulado "An Image Is Worth 16x16 Words (DOSOVITSKIY, 2020): Transformers for Image Recognition at Scale", publicado em 2020 pelos pesquisadores da Google Research. A ideia central por trás do ViT é usar um modelo de transformer para tratar uma imagem como uma sequência de patches. Esses patches são transformados em representações latentes por meio de camadas de transformer, e essas representações são, em seguida, agregadas em uma representação global da imagem. A figura abaixo mostra o diagrama de funcionamento do transformer aplicado à imagens.

Figura 22 - Arquitetura do transformer



Fonte: SZELISKI, 2022

No diagrama vemos que a imagem é dividida em uma série de patches retangulares não sobrepostos, cada um com o mesmo tamanho. Estes patches são então linearmente incorporados em vetores de dimensões fixas. Adicionalmente, a cada *patch* é atribuído um *embedding* de posição, garantindo que o modelo possa considerar a localização espacial original de cada patch na imagem. Esses embeddings são somados aos *embeddings* dos *patches* e juntos alimentam o encoder do transformer. O *encoder* utiliza mecanismos de atenção para ponderar a importância relativa de diferentes *patches*. Em particular, o módulo de *multi-head attention* permite que o modelo considere diferentes combinações de patches simultaneamente, capturando múltiplas representações de interações entre os *patches*. Após o módulo de atenção, os dados passam por uma rede *perceptron* multicamadas (MLP), que realiza transformações adicionais. O processo de atenção e transformação é repetido várias vezes, conforme definido pelo número de camadas do encoder. Especialmente, o mecanismo de *multi-head attention* é fundamental para permitir que o modelo capture interações complexas e múltiplos contextos espaciais entre os patches, oferecendo uma visão rica e holística da imagem.

Para entender o "*multi-head attention*", pense em "cabeças de atenção" como diferentes conjuntos de "lentes" ou "perspectivas" através das quais o modelo visualiza os dados. Cada "cabeça" pode se concentrar em características distintas ou partes da entrada. Por exemplo, uma cabeça pode se concentrar na relação entre o olho e o nariz em um rosto, enquanto outra pode se concentrar na relação entre a boca e o queixo. Ao usar múltiplas cabeças, o modelo pode capturar uma gama mais ampla e rica de interações na imagem.

A operação principal por trás da atenção é a computação de pesos de atenção, que são essencialmente medidas da importância relativa entre diferentes partes da entrada. Estes pesos são calculados usando três componentes principais: consultas (queries), chaves (keys) e valores (values). Em termos simples, as "consultas" e as "chaves" são usadas para determinar os pesos de atenção, e os "valores" são então ponderados por esses pesos para produzir a saída do mecanismo de atenção.

A abordagem ViT tem sido avaliada em vários conjuntos de dados de referência em visão computacional, incluindo o ImageNet, um conjunto de dados de classificação de imagens de grande escala. Os resultados mostram que o ViT supera as arquiteturas de CNNs tradicionais.

Em resumo, o ViT é uma abordagem promissora para tarefas de visão computacional, especialmente para tarefas de classificação de imagens de grande escala. A capacidade do ViT de processar imagens como sequências de patches por meio de camadas de transformer possibilitou melhorias significativas no desempenho e eficiência de tarefas de visão computacional.

3 Trabalhos Relacionados

Neste capítulo serão apresentados e tecidos comentários de alguns trabalhos relacionados ao tema abordado nesta pesquisa. Esses trabalhos serviram de inspiração e ponto de partida para a realização da presente pesquisa. Além disso, os trabalhos mostram a pertinência do tema abordado.

3.1 Facial Key Points Detection using Deep Convolutional Neural Network - NaimishNet

Um problema importante e desafiador na visão computacional é a detecção de pontos chaves na face. Esse problema consiste em identificar as coordenadas x e y de pontos determinados na face humana utilizando para isso uma imagem. Esses pontos são, por exemplo, lateral dos olhos, centro dos olhos, ponta do nariz, etc. São utilizados em aplicações de rastreamento de faces em imagens e vídeos, para reconhecimento de expressões faciais e para detecção de dismorfismo facial em exames médicos, etc. As características faciais variadas entre indivíduos e as diferentes posições das pessoas nas imagens e as variadas condições de iluminação tornam essa tarefa bem desafiadora (NAIMISH, 2017).

Vale dizer que os trabalhos que se propõem a identificar tais pontos utilizam majoritariamente redes neurais convolucionais, alguns deles aplicaram algumas técnicas de processamento da imagem como alongamento de histograma antes de aplicar as redes neurais (NAIMISH, 2017).

Em 2017, Naiminsh escreveu o trabalho “*Facial Key Points Detection using Deep Convolutional Neural Network*”. Nele, a rede neural NamishNet foi proposta, e se consiste em uma variação da LeNet composta por 4 camadas convolucionais 2D, 4 camadas de agrupamento máximo 2D e 3 camadas densas, intercaladas com camadas de descarte e funções de ativação. A tabela abaixo mostra a arquitetura de rede utilizada.

Tabela 1 - Arquitetura NamishNet

Layer Name	Number of Filters	Filter Shape
$Convolution2d_1$	32	(4, 4)
$Convolution2d_2$	64	(3, 3)
$Convolution2d_3$	128	(2, 2)
$Convolution2d_4$	256	(1, 1)

Fonte: CFAZ.NET, 2020

Foram avaliadas muitas arquiteturas de redes antes de se chegar à arquitetura final adotada, não sendo feitos experimentos utilizando zero padding e strides. As decisões foram tomadas levando em consideração as restrições do hardware disponível e tempo necessário para treinamento. Eles treinaram um modelo diferente para prever cada um dos pontos chaves na face.

O treinamento da rede foi feito utilizando banco de dados de pontos faciais disponível na Kaggle Competition – Facial Key Points Detection. Os resultados da competição no kaggle foram utilizados para balizar a eficiência do modelo proposto. O bando de dados conta com 7049 imagens nas quais são identificadas as coordenadas de 15 pontos na face. O banco de teste consiste de 1783 imagens sem informações dos pontos.

A pontuação dos modelos no kaggle foi feita utilizando a predição dos pontos em algumas imagens e computando o erro médio quadrático (RMSE).

$$RSME = \sqrt{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \div n}$$

Na fórmula y_i é o valor original do ponto e $\hat{y}_i$ É o valor da predição do ponto. O somatório de todos os erros permite comparar qual modelo tem maior precisão para encontrar os pontos.

Durante o treinamento, percebeu-se que aumentar o número de dados invertendo todas as imagens melhorou a precisão do modelo. Também verificou-se que normalizar os valores de entrada e centralizar em 0 melhorou a precisão do modelo.

Ao fim do trabalho foi constatado que o modelo baseado na rede LeNet obteve sucesso ao prever os pontos faciais em uma imagem da face. Também foi dada atenção à possibilidade de alguns trabalhos futuros que incluem: fazer experimentos com maior poder computacional, utilizar outras arquiteturas de redes como a VGGNet, e fazer a detecção em etapas fazendo primeiro a detecção da face e depois o alinhamento da face antes de prever os pontos (NAIMISH, 2017).

3.2 Mask R-CNN

Em 2017 Kaiming He propôs um framework para segmentação de objetos. A sua abordagem identifica objetos na imagem ao mesmo tempo que cria uma máscara de segmentação para cada um dos objetos. Esse framework ficou conhecido como Mask R-CNN e é uma extensão das Faster R-CNN também proposta por He. A diferença é que agora foi adicionado um passo para prever a máscara do objeto juntamente com a caixa de contenção do objeto. O Mask R-CNN é mais simples de treinar do que a Faster R-CNN e também são mais facilmente adaptáveis para realizar outras tarefas, como por exemplo estimar poses humanas com o mesmo framework (He et al., 2017).

O framework foi utilizado para realizar os desafios do banco de dados COCO - Common Objects in Context (LIN, 2015). Ele obteve melhores resultados nos desafios de segmentação de imagem, detecção da caixa de contenção de objetos e detecção de pontos em poses humanas. O código do framework foi disponibilizado em https://github.com/facebookresearch/Detectron para que outras pessoas possam utilizá-lo como base para futuras pesquisas.

Segundo He, já existem boas técnicas e ferramentas para realizar a classificação de imagens. O objetivo deste trabalho foi produzir um framework para segmentação de objetos tão bom quanto os já existentes para classificação de imagens. A segmentação de objetos é desafiadora pois ela exige a identificação correta dos objetos na imagem bem como realizar a segmentação de cada objeto. Essa tarefa exige combinar duas áreas da visão computacional. A detecção de objetos para encontrar a caixa de seleção dos objetos. E também a semântica de objetos para identificar cada pixel que compõe a máscara do objeto. No seu trabalho ele criou um framework surpreendentemente simples capaz de realizar essa tarefa.

A Mask R-CNN é uma extensão da Faster R-CNN. Uma das principais melhorias é que agora foi realizado o alinhamento dos pixels de saída com os pixels de entrada do modelo. Isso foi feito realizando um passo a mais de alinhamento das regiões de interesse ROI. Essa mudança melhorou a precisão da máscara de 10% para 50%. Os modelos levam aproximadamente 200ms por imagem em uma GPU. O treinamento de todo o banco de dados COCO levou 2 dias em uma única máquina com 8 GPUs.

Mask R-CNN são conceitualmente simples. A Faster R-CNN apresenta duas saídas para cada objeto. Um caixa que determina a região do objeto e um label que identifica o objeto. A Mask R-CNN simplesmente adiciona uma nova saída com a segmentação dos pixels do objeto. O ponto chave para a Mask R-CNN é o alinhamento dos pixels da saída com os pixels de entrada. O alinhamento dos pixels na máscara com as imagens de entrada foi possível porque cada máscara proposta tem o tamanho exato da caixa de contenção associada. Isso permitiu fazer o alinhamento de cada pixel da máscara com a imagem.

Faster R-CNN consiste em dois estágios. O primeiro estágio é chamado de Region Proposal Network (RPN), que propõe caixas de contenção para objetos. O segundo estágio é a Fast CNN, que extrai características de cada uma das caixas de contenção para fazer a classificação do objeto e regressão das caixas de contenção. A Mask R-CNN tem o mesmo primeiro estágio mas também identifica uma máscara de pixels para cada caixa de contenção proposta. No segundo estágio foi acrescentado à saída a máscara de segmentação junto à caixa de contenção e a classificação.

A arquitetura do framework Mask R-CNN é flexível, então variações da arquitetura podem ser feitas. Por padrão a arquitetura é formada por 3 blocos:

1. **Backbone Network:** Extrai mapas de características da imagem em diferentes escalas. Essa etapa utiliza uma ResNet de profundidade 50 ou 101.
2. **Region Proposal Network:** Detecta regiões que podem conter os objetos nas diferentes escalas obtidas no backbone. Por padrão são encontradas 1000 regiões com os respectivos níveis de confiança.
3. **Box Head:** Refina a posição da caixa de contenção utilizando camadas completamente conectadas.

Em seu artigo, He utilizou o framework Mask R-CNN para prever pontos chaves em poses humanas. Os pontos chaves são ombro direito, joelho esquerdo, etc. Para isso foi realizada uma alteração no sistema de segmentação do framework, na qual, para cada objeto encontrado foi encontrada uma máscara de segmentação para cada ponto. A máscara tem o tamanho do objeto e apenas um pixel destacado para indicar cada ponto. O treinamento foi realizado no banco de dados COCO de poses humanas. Como haviam poucas imagens, novas imagens foram criadas utilizando variações de tamanho nas imagens originais. A imagem mostra o resultado dos pontos de pose encontrados com o framework Mask R-CNN.

Figura 23 - Representação de pontos chaves em formas humanas

Fonte: HE et al., 2017

Nos apêndices do artigo foram realizados outros experimentos aplicando o framework no banco de dados Cityscapes Dataset – Semantic Understanding of Urban Street Scenes (DAIMLER, 2016). Esse banco de dados contém 2975 imagens de treinamento, 500 imagens de validação e 1525 imagens para teste. Um dos desafios de se trabalhar com esse banco de dados é a quantidade pequena de dados para treinamento, então foi utilizado um modelo pré-treinado com o banco de dados COCO. Essa abordagem mostrou que usar modelos pré-treinados melhoraram o resultado nesse banco de dados (He et al., 2017).

3.3 ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation

Em 2022, Xu et al. introduziram o artigo intitulado "ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation" (XU et al., 2022). Este trabalho propôs uma adaptação do modelo Vision Transformer específica para a detecção de pontos de pose em figuras humanas. Em 2023, este artigo é reconhecido como o estado da arte na detecção de pontos em humanos, utilizando a base de dados COCO - Common Objects in Context (LIN, 2015).

Assim como em todos os modelos baseados em "transformers", a imagem inicial é segmentada em patches de tamanho fixo. Um encoder é empregado para adicionar informações de posição a esses patches. No decorrer do encoder, diversas camadas de transformadores são empregadas para gerar tokens que representam a imagem. Cada uma dessas camadas é constituída por uma sequência específica de componentes: inicia-se com uma camada de normalização, seguida por uma camada de autoatenção com múltiplas cabeças e uma conexão de salto (skip connection). Posteriormente, uma segunda camada de normalização é aplicada, seguida de uma rede neural feed-forward simples, culminando em outra conexão de salto. A saída desse mecanismo é direcionada a um decoder, para o qual foram propostas duas alternativas.

A primeira, denominada "decoder clássico", é composta por dois blocos de desconvolução, responsáveis por aumentar a resolução espacial do mapa de características. Cada bloco contém uma camada de desconvolução seguida de normalização em lote e a função de ativação ReLU. A saída desses blocos é processada por um preditor que elabora um mapa de calor para cada ponto-chave identificado na imagem. Por exemplo, na detecção de 17 pontos de pose humana, são gerados 17 mapas de calor, cada um representando a localização de um ponto específico na imagem.

Já a segunda alternativa, nomeada "decoder simples", amplia diretamente a dimensão espacial do mapa de características em quatro vezes através da interpolação bilinear. Este processo é seguido pela função de ativação ReLU e, posteriormente, por uma camada de convolução com um kernel de tamanho 3 × 3, responsável por produzir os mapas de calor. Embora o "decoder simples" apresente uma capacidade de não

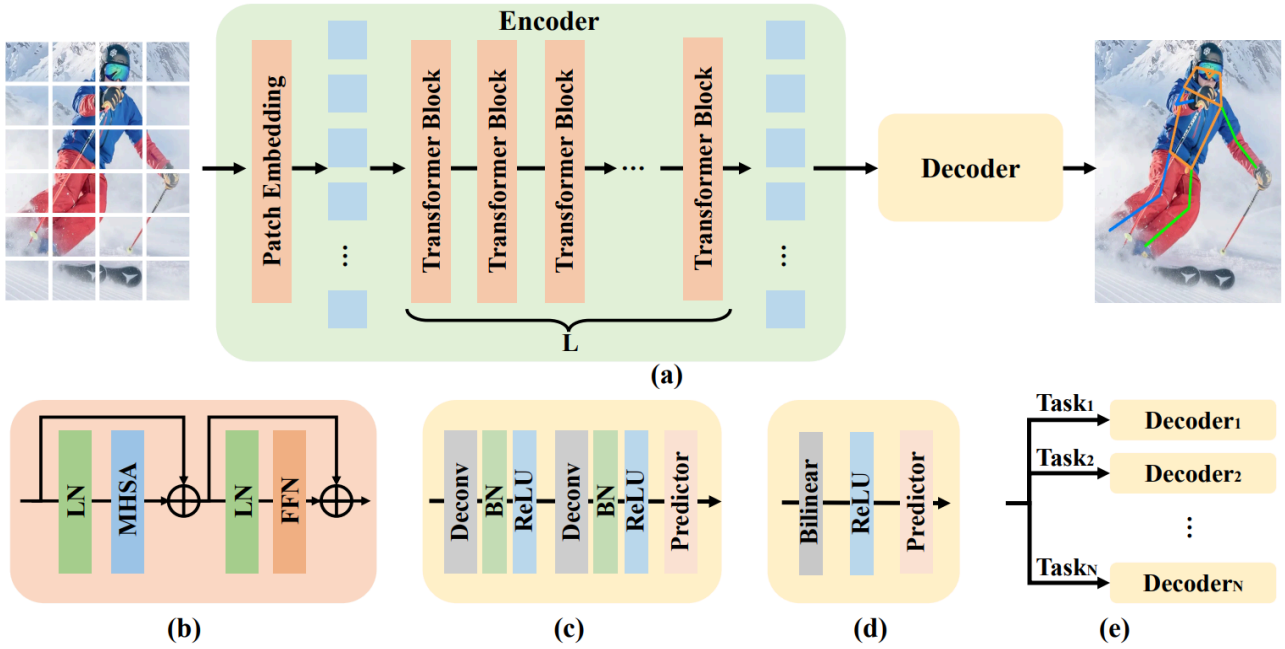
linearidade inferior em relação ao "decoder clássico", ele demonstrou uma performance comparável quando avaliado lado a lado com o primeiro.

O mapa de calor gerado é utilizado para estimar a localização dos pontos de pose nas imagens.

A proposta central do artigo é a apresentada acima, no entanto, o trabalho também inova ao sugerir o treinamento utilizando múltiplas bases de dados, cada qual com diferentes quantidades de pontos de pose como, por exemplo, uma base com 17 pontos e outra com 25 pontos. Para atender a essa variedade, mantém-se um encoder padrão, mas incorporam-se diversos decoders, com cada um projetado para prever uma quantidade distinta de pontos. A seleção do decoder adequado durante o treinamento é feita por meio de uma condicional que identifica a base de dados da imagem em processamento. Essa estratégia possibilitou uma melhoria na performance do modelo, uma vez que ampliou a quantidade de dados para treinamento em comparação ao uso de uma única base de dados.

A Figura 24 mostra as arquiteturas discutidas no artigo: (a) Estrutura do ViTPose. (b) Bloco transformer. (c) Decodificador clássico. (d) Decodificador simples. (e) Decodificadores para variados conjuntos de dados.

Figura 24 - Estruturas do VitPose



Fonte: XU et al., 2022

Neste artigo, foram propostos quatro diferentes modelos, nomeados como: modelo base, modelo grande, modelo enorme e modelo gigante. A principal distinção entre esses modelos reside no número de camadas e parâmetros, que incrementa progressivamente de um modelo para o outro. Para avaliar a eficiência dos modelos propostos, os autores compararam-nos com estruturas de identificação de pontos que empregam redes neurais baseadas em ResNet-50 e ResNet-152 como encoders. A análise revelou que o uso do vision transformer como encoder superou em desempenho as estruturas baseadas em ResNet, independentemente de se usar o decodificador clássico ou o simples. A Figura 25 apresenta um comparativo de precisão entre os modelos avaliados.

Figura 25 - Comparativo da estrutura do VitPose

Backbone	ResNet-50		ResNet-152		ViTPose-B		ViTPose-L		ViTPose-H	
Decoder	Classic	Simple	Classic	Simple	Classic	Simple	Classic	Simple	Classic	Simple
AP	71.8	53.1	73.5	55.3	75.8	75.5	78.3	78.2	79.1	78.9
AP_{50}	89.8	86.9	90.5	87.9	90.7	90.6	91.4	91.4	91.7	91.6
AR	77.3	62.0	79.0	63.8	81.1	80.9	83.5	83.4	84.1	84.0
AR_{50}	93.7	92.1	94.3	92.9	94.6	94.6	95.3	95.3	95.4	95.4

Fonte: XU et al., 2022

O artigo ressalta uma observação crucial sobre a eficácia dos decodificadores. Quando se substituiu o decodificador clássico pelo simples, em uma estrutura que utilizava a ResNet como encoder, houve uma queda significativa no desempenho. Entretanto, essa drástica redução de eficiência não foi observada ao utilizar o vision transformer como encoder, onde a diminuição da performance foi apenas marginal.

Além disso, o estudo também avaliou a influência do volume de dados utilizados no pré-treinamento do modelo. O pré-treinamento envolve inicialmente o treino do modelo em um conjunto de dados extenso para posteriormente realizar ajustes mais específicos em um conjunto de dados alvo. O experimento avaliou a performance do modelo utilizando um conjunto inicial de 1 milhão de imagens da base ImageNet. Posteriormente, a comparação de desempenho foi realizada com subconjuntos de 500 mil, 300 mil e 150 mil imagens. Notavelmente, o artigo revelou que o modelo manteve um desempenho consistente mesmo quando treinado com um volume menor de imagens. A Figura 26 ilustra os resultados dessa comparação.

Figura 26 - Avaliação do pré treinamento no VitPose

Pre-training Dataset	Dataset Volume	AP	AP_{50}	AP_{75}	AR	AR_{50}	AR_{75}
ImageNet-1k	1M	75.8	90.7	83.2	81.1	94.6	87.7
COCO (cropping)	150K	74.5	90.5	81.9	80.0	94.5	86.6
COCO+AI Challenger (cropping)	500K	75.8	90.8	83.0	81.0	94.6	87.4
COCO+AI Challenger (no cropping)	300K	75.8	90.5	83.0	81.0	94.5	87.4

Fonte: XU et al., 2022

O estudo também investigou o impacto da resolução das imagens na performance do modelo, considerando que a resolução de uma imagem pode afetar a quantidade de detalhes disponíveis para análise e, consequentemente, a precisão da detecção de pontos de pose. Ao variar a resolução das imagens utilizadas no treinamento e teste, o artigo buscou determinar a relação entre qualidade da imagem e eficiência do modelo. A Figura 27 apresenta os resultados obtidos nessa avaliação.

Figura 27 - Avaliação da resolução no VitPose

	224x224	256x192	256x256	384x288	384x384	576x432
AP	74.9	75.8	75.8	76.9	77.1	77.8
AR	80.4	81.1	81.1	81.9	82.0	82.6

Fonte: XU et al., 2022

Ao analisar os resultados, notamos uma distinção significativa entre a primeira e a segunda coluna: mesmo a primeira coluna apresentando uma maior quantidade de pixels, a alteração das dimensões da imagem na segunda coluna produziu resultados superiores. Isso pode ser justificado pela característica intrínseca das imagens analisadas, que retratam figuras humanas e, portanto, possuem uma dimensão vertical mais acentuada que a horizontal. Ao examinar as demais colunas, identifica-se que um aumento progressivo na resolução contribui para uma melhoria sutil nos resultados.

O estudo também avaliou o incremento na performance ao se empregar múltiplos conjuntos de dados, em contraste com a utilização de um único conjunto. Foi constatado que a integração de diversas bases de dados potencializa ligeiramente a eficácia do modelo. Os dados que corroboram essa afirmação estão representados na Figura 28.

Figura 28 - Avaliação de múltiplas base de dados no VitPose

COCO	AIC	MPII	AP	AP_{50}	AR	AR_{50}
✓			75.8	90.7	81.1	94.6
✓	✓		77.0	90.8	82.2	94.9
✓	✓	✓	77.1	90.8	82.2	94.7

Fonte: XU et al., 2022

Em conclusão, o artigo coloca lado a lado o desempenho do modelo em relação a outras arquiteturas de ponta dedicadas à tarefa de identificação de pontos de pose na base de dados COCO. A Figura 29 exibe os resultados alcançados. É relevante ressaltar que os modelos marcados com um asterisco (*) denotam aqueles que foram submetidos a treinamentos utilizando múltiplos conjuntos de dados.

Figura 29 - Comparativo do estado da arte com VitPose

Model	Backbone	Params (M)	Speed (fps)	Input Resolution	Feature Resolution	COCO val AP	COCO val AR
SimpleBaseline [42]	ResNet-152	60	829	256x192	1/32	73.5	79.0
HRNet [36]	HRNet-W32	29	916	256x192	1/4	74.4	78.9
HRNet [36]	HRNet-W32	29	428	384x288	1/4	75.8	81.0
HRNet [36]	HRNet-W48	64	649	256x192	1/4	75.1	80.4
HRNet [36]	HRNet-W48	64	309	384x288	1/4	76.3	81.2
UDP [18]	HRNet-W48	64	309	384x288	1/4	77.2	82.0
TokenPose-L/D24 [27]	HRNet-W48	28	602	256x192	1/4	75.8	80.9
TransPose-H/A6 [44]	HRNet-W48	18	309	256x192	1/4	75.8	80.8
HRFormer-B [48]	HRFormer-B	43	158	256x192	1/4	75.6	80.8
HRFormer-B [48]	HRFormer-B	43	78	384x288	1/4	77.2	82.0
ViTPose-B	ViT-B	86	944	256x192	1/16	75.8	81.1
ViTPose-B*	ViT-B	86	944	256x192	1/16	77.1	82.2
ViTPose-L	ViT-L	307	411	256x192	1/16	78.3	83.5
ViTPose-L*	ViT-L	307	411	256x192	1/16	78.7	83.8
ViTPose-H	ViT-H	632	241	256x192	1/16	79.1	84.1
ViTPose-H*	ViT-H	632	241	256x192	1/16	79.5	84.5

Fonte: XU et al., 2022

Ao analisar os dados apresentados, é evidente que o modelo ViTPose superou em precisão outros modelos considerados referências no estado da arte. Notavelmente, além de sua superioridade em acurácia, o ViTPose também se destaca pela sua rapidez em predição, conforme demonstrado na coluna de FPS (frames por segundo). Este aspecto é

particularmente relevante, uma vez que indica que, apesar de sua elevada precisão, o modelo não compromete sua velocidade em comparação com outros modelos existentes.

3.4 'Aariz: A Benchmark Dataset for Automatic Cephalometric Landmark Detection and CVM Stage Classification

No ano de 2023, pesquisadores do Paquistão sugeriram a constituição de uma base de dados compreendendo telerradiografias e marcos anatômicos com um total de 1000 imagens, almejando estabelecê-la como referência na formação de bancos de dados para treinamento de modelos. Segundo o artigo, as ferramentas existentes para detecção automática de pontos em telerradiografias revelam-se insatisfatórias para aplicação ortodôntica, primordialmente devido à carência de um vasto banco de telerradiografias apropriado ao treinamento desses modelos. Este repositório proposto inclui imagens provenientes de sete diferentes dispositivos de aquisição de radiografia, cada qual com suas resoluções específicas, consolidando-se assim como a base de dados mais diversificada e abrangente disponível até aquele momento. Os autores demarcaram 29 marcos anatômicos nas imagens, contando com a expertise dos especialistas das instituições acadêmicas associadas ao projeto. Adicionalmente, anotaram o estágio de maturação da coluna cervical em todas as radiografias. Desta forma, além de sua aplicação primária, essa base de dados também se apresenta como recurso para investigações relacionadas à maturação vertebral cervical, estabelecendo-se como o primeiro conjunto de dados apropriado à classificação da maturidade cervical (KHALID et al., 2023).

Na introdução, os autores destacam a importância da morfometria do crânio e análises cefalométricas para a ortodontia e cirurgias orais. Estas análises são feitas em cefalogramas, onde a marcação manual de pontos é complexa e varia entre especialistas. Apesar dos avanços em sistemas automáticos, sua eficácia é limitada pela falta de dados adequados. Enfatizam a relevância de avaliações precisas do crescimento facial para tratamentos ortodônticos. E então sugerem um conjunto de dados inovador, 'Aariz, com 1000 imagens cefalométricas, visando aprimorar essa área.

Na sessão de trabalhos relacionados os autores citam alguns trabalhos que utilizam aprendizagem de máquina e visão computacional para realizar a detecção de

pontos anatômicos em telerradiografias. Apresentam alguns conjuntos de dados que foram criados para estes estudos e ressaltam que eles têm limitações, como tamanho, abrangência ou falta de disponibilidade pública. Com isso, eles buscam justificar que a criação da base de dados proposta pelo artigo é relevante.

O artigo descreve a organização da estrutura de dados empregada na constituição da base de dados. Foram estabelecidos pontos anatômicos, para os quais foi elaborada uma tabela referencial. Essa tabela categoriza cada ponto com um número, um nome e uma identificação relativa à sua localização: estrutura óssea, arcada dentária ou tecido mole. Adicionalmente, foi desenvolvida uma tabela de referência para detalhar as especificações dos equipamentos utilizados na aquisição das imagens. Essa tabela informa características de cada aparelho, incluindo o fabricante, a densidade de pixels por milímetro e a quantidade de imagens obtidas por cada dispositivo. Quanto à demarcação dos pontos nas telerradiografias, o processo ocorreu em duas etapas. Na primeira, dois ortodontistas juniores, com cinco anos de experiência, marcaram os pontos. Na segunda etapa, a marcação foi realizada por dois ortodontistas seniores com uma década de experiência. O valor referencial para cada ponto foi estabelecido pela média dessas duas demarcações.

Por fim, na conclusão, os autores reiteram que a falta de conjuntos de dados confiáveis tem limitado o desenvolvimento de sistemas automáticos de detecção de marcos. Portanto, para preencher essa lacuna, apresentaram um novo conjunto com radiografias cefalométricas laterais anotadas com 29 pontos anatômicos, totalizando 1000 radiografias de 7 dispositivos diferentes. Além das anotações, especialistas classificaram o estágio de maturação vertebral cervical (CVM) de cada imagem. Com sua diversidade e abrangência, seu conjunto pode aprimorar a precisão dos sistemas de detecção de pontos cefalométricos, contribuindo para decisões ortodônticas mais informadas.

3.5 Contextualização de trabalhos relacionados

Os trabalhos detalhados neste capítulo foram escolhidos por apresentarem uma alta correlação com a presente pesquisa. Foram selecionados três artigos que descrevem técnicas de visão computacional recentes, e ambos comprovaram sua eficácia em aplicações semelhantes ao objetivo da presente pesquisa. Também foi selecionado um

artigo que propõe a criação de uma base de dados similar a que foi criada para realização dessa pesquisa.

O artigo descrito no capítulo 3.1 apresenta a utilização de uma rede neural convolucional para detectar pontos chave em uma fotografia de rosto humano. Essa técnica pode ser aplicada de forma similar para a detecção de pontos anatômicos em radiografias em norma lateral.

No artigo descrito no capítulo 3.2, é apresentada uma aplicação da R-CNN para detecção de pontos em articulações do corpo em fotografias de pessoas com a finalidade de descrever sua pose. O artigo demonstra que redes neurais convolucionais regionais podem ser eficientemente utilizadas para prever essas poses. Uma técnica semelhante pode ser aplicada na detecção de pontos anatômicos em radiografias em norma lateral.

No capítulo 3.3, é apresentado um artigo que emprega a arquitetura de Transformers para detecção de pontos articulares em fotografias humanas, objetivando a descrição precisa da pose. Esta abordagem evidencia como as inovações recentes relacionadas aos transformadores têm sido adaptadas para processamento de imagens. A eficácia demonstrada nesse contexto, que guarda semelhanças com o desafio de identificar pontos anatômicos em telerradiografias, sugere que a arquitetura de Transformers pode ser considerada uma estratégia promissora para a problemática investigada neste trabalho de pesquisa.

No capítulo 3.4, é apresentado um artigo que propõe a elaboração de uma base de dados de telerradiografias para servir de referência para realização de pesquisas relacionadas a pontos anatômicos localizados em telerradiografias laterais. Como o conjunto de dados desenvolvido para esta pesquisa supera em magnitude o apresentado na referida publicação, isso destaca a relevância do mesmo.

Os estudos em referência demonstram técnicas de visão computacional eficazes em aplicações análogas à detecção de pontos anatômicos em telerradiografias laterais. Essa eficácia sugere que estas técnicas são promissoras para essa finalidade e, consequentemente, merecem ser avaliadas neste estudo. Especificamente no contexto do exame de cefalometria, observa-se uma lacuna em soluções satisfatórias para aplicação ortodôntica e uma carência de bases de dados diversificadas e abrangentes, conforme evidenciado por Khalid et al. (2023).

4 Metodologia

O trabalho foi desenvolvido em cinco etapas, uma para cada um dos objetivos específicos a serem alcançados e mais uma para finalização da escrita do projeto final. As etapas serão: avaliação do estado da arte, construção da base de conhecimento, classificação de técnicas mais adequadas, criação do traçado cefalométrico para referência comparativa e finalização da escrita da dissertação.

1. **Avaliação do estado da arte:** A primeira etapa consiste na pesquisa extensiva de técnicas de visão computacional disponíveis na literatura. Foi realizada através da leitura de livros e artigos acadêmicos sobre o assunto, além de consulta aos professores do CEFET-MG que atuam em áreas de estudo correlatas. Nessa etapa do trabalho, levantou-se quais técnicas de visão computacional existem e dentre essas quais são candidatas promissoras à tarefa de encontrar pontos cefalométricos em telerradiografias. Depois de levantadas essas técnicas obteve-se uma lista de técnicas que foram avaliadas na terceira etapa do trabalho. Durante a etapa da pesquisa foram produzidos artigos sobre tópicos específicos que foram considerados interessantes.

2. **Construção da base de conhecimento:** Na segunda etapa da pesquisa foi construída uma base de conhecimento a partir de dados de exames de traçado cefalométrico feitos por especialistas. A base de dados utilizada foi obtida a partir dos exames realizados ao longo de 8 anos na plataforma SaaS Cfaz.net, empresa fundada pelo pesquisador autor deste trabalho. O Cfaz.net é um sistema online para centros médicos e odontológicos largamente utilizado no Brasil e América Latina. O sistema oferece, dentre outras funcionalidades, uma ferramenta de auxílio ao diagnóstico para confecção do traçado computadorizado. Nessa ferramenta, o especialista precisa marcar os pontos cefalométricos manualmente para que o sistema realize os cálculos necessários para confecção do exame. Por ser um sistema SaaS os dados dos exames estão concentrados em um mesmo

banco de dados, o que permite a extração de uma base de conhecimento grande e relevante.

3. **Escolha de técnicas mais adequadas:** A terceira etapa do trabalho consistiu em usar a base de conhecimento adquirida na segunda etapa para elencar as técnicas de visão computacional mais adequadas que foram avaliadas neste trabalho. A base de conhecimento foi dividida em uma base de treinamento, uma base de validação e uma base de testes. A base de treinamento e validação foi utilizada para desenvolvimento e escolha de hiperparâmetros dos modelos matemáticos computacionais. A base de testes foi usada para validar cada um dos modelos criados e garantir que eles sejam capazes de realizar a tarefa a que se propõem.

4. **Criação do traçado cefalométrico para referência comparativa:** Na quarta etapa do projeto, vários especialistas realizaram o exame de traçado cefalométrico na mesma telerradiografia. As marcações de pontos desses especialistas foram usadas para criar uma referência padrão, elaborada a partir da média dos pontos anatômicos marcados. Dessa forma, temos uma referência confiável para a marcação correta dos pontos. Essa referência foi utilizada para criar um quadro comparativo do erro acumulado de cada especialista, bem como de cada modelo de marcação de pontos criado no projeto. O objetivo desse quadro é mostrar aos profissionais de radiologia, que não têm conhecimento em computação, redes neurais, base de dados e minimização de erros médios quadráticos, a capacidade dos modelos desenvolvidos.

5. **Finalização da escrita da dissertação:** A quinta etapa foi dedicada à escrita da versão final da dissertação. Nessa etapa, o texto produzido foi apresentado a terceiros para que sejam feitas sugestões e erros sejam apontados. Os erros foram corrigidos e as sugestões levadas em consideração para elaboração da versão final da dissertação.

4.1 Coleta e estruturação de dados

Um dos pontos-chaves para desenvolver algoritmos de visão computacional de maneira confiável é utilizar uma base de dados desafiadora e representativa (SZELISKI, 2022). As técnicas de visão computacional são altamente dependentes de dados para extrair o conhecimento sobre como realizar uma tarefa. Esses dados precisam ser estruturados de uma maneira que os algoritmos possam processá-los e a precisão do modelo treinado depende da quantidade e qualidade da base de treinamento que foi utilizada durante sua criação.

Não existe uma regra clara para saber a quantidade de dados necessária para treinar um modelo. Apesar da literatura citar uma regra geral que recomenda utilizar mil exemplos para classe em um problema de classificação, não existe uma recomendação similar para todos os problemas. A quantidade de dados necessária depende da complexidade do problema a ser resolvido e também da complexidade do algoritmo utilizado. O tamanho ideal da base de dados precisa ser avaliada empiricamente assim como a maioria dos hiperparâmetros de uma rede neural. Um bom ponto de partida é verificar trabalhos relacionados e observar o tamanho da base de dados que foi utilizada. Por exemplo, no trabalho de detecção de pontos na face vimos que as Redes Neurais Convolucionais foram utilizadas em uma base de dados de pouco mais de 7 mil imagens com resultados satisfatórios (NAIMISH, 2017).

Nos trabalhos utilizando transformers para classificação de imagens foram utilizados uma base de dados muito superior, com mais de 300 milhões de exemplos (DOSOVITSKIY, 2020). Entretanto, o trabalho realizado por Dosovitskiy classificou milhares de classes em milhões de imagens extremamente variadas. Para a tarefa de marcar pontos em radiografias não temos tanta variedade de imagens como entrada nem tantas opções para saída. Sendo assim imagino que não será necessário uma base de dados tão grande.

O único banco de dados público de pontos cefalométricos encontrado está disponível no kaggle e conta com apenas 400 imagens. Baseado nos trabalhos relacionados consideramos esse número insuficiente. Além disso, as imagens eram todas de um mesmo aparelho de raios-x, o que não traria a representatividade ideal para a base de dados.

Para essa pesquisa construímos nossa própria base de dados obtido através dos exames realizados no sistema de auxílio ao diagnóstico Cfaz.net. O sistema conta com um software para realização do exame de traçado cefalométrico em uso desde 2014.

Foram avaliadas as questões legais para utilização desses dados. Verificamos que os usuários da plataforma aceitaram a política de privacidade do sistema que afirma que os dados podem ser usados para desenvolver o processo de realização de exames. Além disso, existe uma legislação específica para definir as diretrizes legais referentes ao tratamento de dados pessoais, a lei geral de proteção de dados Nº 13.709 em vigor desde 2018. Segundo essa lei, dados podem ser usados para fins de pesquisa desde que sejam anonimizados, ou seja, não contenham informações que permitam identificar os usuários (BRASIL, 2018).

Para construção da base de dados extraímos um conjunto da totalidade de pontos cefalométricos da análise da USP. Essa análise foi escolhida porque é a mais utilizada no Brasil, sendo requisitada em 34% dos exames realizados. Foram utilizados apenas os 26 pontos necessários para desenhar as linhas e planos cefalométricos e gerar a tabela de fatores:

1. N - Násio
2. Or - Orbital
3. S - Sela
4. Po - Pório
5. Go - Gônio
6. Me - Mentoniano
7. Pog - Pogônio
8. Gn - Gnátio
9. E - Ponto E
10. B - Ponto B
11. D - Ponto D
12. A - Ponto A
13. P' - Ponto P Linha
14. Ppd - Ponto Posterior de Downs
15. Aii - Ápice Incisivo Inferior

16. Iii - Incisal do Incisivo Inferior
17. Iis - Incisal do Incisivo Superior
18. Sf1/ - Face vestibular do incisivo central superior
19. Ais - Ápice Incisivo Superior
20. Pog' - Pogônio do Tecido Mole
21. Ls - Lábio Superior
22. Pn - Pronasal
23. V - Ponto V
24. T - Ponto T
25. Tuber - Tuber
26. Pi - Protuberância Incisal

Além destes, ainda existem pontos marcados no software para desenhar estruturas que não impactam nos valores do exame e não são utilizados para o diagnóstico, são apenas um recurso estético para facilitar a visualização das estruturas da face. É comum a remoção de estruturas nas configurações do software e com isso alguns exames não tem esses pontos. Então, para evitar dados nulos na nossa base de dados, optamos por utilizar apenas os 26 pontos. Para a construção do banco de dados foram utilizados 30 mil exames de traçado cefalométrico com a análise cefalométrica no padrão USP.

Para realizar a extração e transformação dos dados, implementamos um *script* em Ruby para acessar o banco de dados do sistema. Os valores levantados foram organizados em um arquivo CSV no qual cada ponto demarcado por cada especialista foi representado por dois valores: a posição na imagem em pixels para o eixo X e a posição na imagem em pixels para o eixo Y. Além disso, armazenamos o id do exame no Cfaz.net, a altura da imagem em pixels, a largura da imagem em pixels e um link para a imagem armazenada em um storage.

O formato da tabela pode ser observado na Figura 30:

Figura 30 - Visualização da base de dados

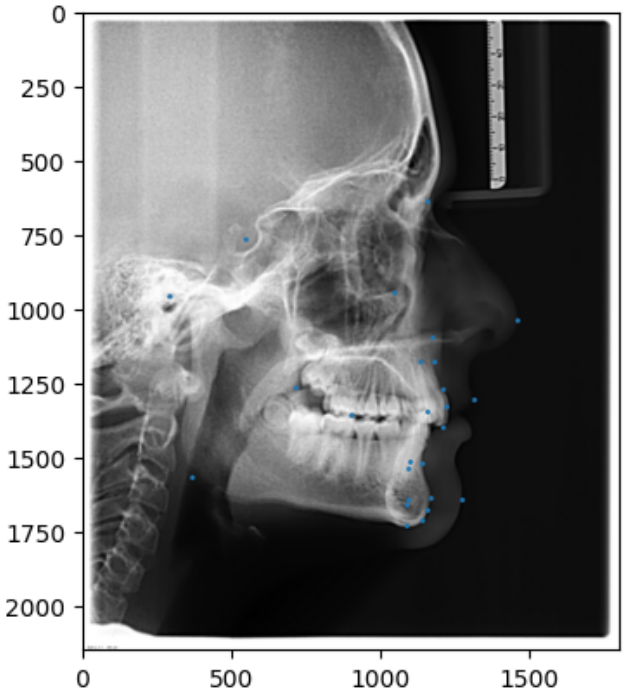
	2_x	2_y	3_x	3_y	7_x	7_y	11_x	11_y	23_x	23_y	...	90_x	90_y	91_x	91_y	92_x
0	1263.58	483.810	1188.00	770.256	639.094	646.925	448.169	841.867	627.161	1422.72	...	1275.51	1426.69	853.884	1120.36	1378.93
1	1159.01	633.178	1048.39	942.490	548.255	764.492	292.988	953.060	366.129	1562.82	...	1095.98	1509.81	716.387	1262.98	1208.72
2	1469.83	958.744	1294.58	1245.580	640.598	1035.800	358.490	1258.420	614.952	2003.34	...	1439.91	2007.62	858.591	1648.01	1491.20
3	1405.98	1052.460	1316.07	1368.640	656.678	1159.280	391.211	1347.280	656.678	1975.37	...	1435.95	1996.73	900.737	1701.92	1534.43
4	1906.39	646.907	1811.72	994.080	1239.210	809.222	1009.310	985.062	1176.100	1647.85	...	1879.34	1719.99	1433.060	1318.71	1942.45
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
29995	1548.22	525.359	1376.67	922.570	698.840	721.874	426.871	935.114	694.656	1671.00	...	1514.75	1737.90	970.809	1353.23	1623.53
29996	1633.65	594.794	1498.57	941.868	850.191	748.048	593.541	932.853	733.122	1676.58	...	1566.11	1726.17	1025.790	1306.97	1656.16
29997	1517.19	531.037	1430.01	882.502	729.849	727.526	436.500	904.642	671.733	1754.25	...	1548.82	1715.91	974.575	1435.01	1659.52
29998	1592.27	638.168	1461.62	975.831	844.438	777.735	583.147	930.809	736.318	1651.16	...	1542.72	1678.17	1024.640	1286.48	1601.28
29999	1462.49	675.168	1319.96	972.668	682.068	806.622	441.301	1034.940	582.440	1740.63	...	1375.31	1711.57	881.323	1353.19	1537.21

30000 rows × 56 columns

Fonte: Produzido pelo autor, 2022

A Figura 31 mostra uma imagem com os pontos marcados para exemplificar a Figura 30:

Figura 31 - Visualização da base de dados



Fonte: Autoria própria, 2022

Em 942 exames havia pelo menos um ponto que não foi marcado, isso pode ter acontecido porque é possível customizar as análises no software e algum especialista pode ter customizado a análise da USP. A customização pode ter removido alguns fatores ou planos o que ocasiona consequentemente na remoção dos pontos. Essas linhas foram descartadas para que tivéssemos apenas exames com todos os pontos para treinamento. Então a base de dados disponível tem 29.058 exames da USP completamente marcados.

Criamos um *script* em Python para baixar cada uma das imagens referenciadas na base de dados e colocar em uma pasta no HD local para realizar o treinamento dos modelos.

Levando em consideração a análise feita anteriormente consideramos essa base de dados suficiente para a realização do trabalho. Durante a realização dos testes precisamos ter atenção para a taxa de erro na base de treinamento e taxa de erro na base de testes. Se a taxa de erro na base de treinamento for significativamente maior que a base de testes estamos tendo *overfit* do modelo e mais dados podem ser necessários. Também existem técnicas na literatura para lidar com o *overfit* que podem ser aplicadas se necessário. Caso as taxas de erros estejam similares, saberemos que nossa base de dados tem tamanho suficiente.

4.2 Técnicas mais adequadas

Analisando a história da visão computacional podemos ver que inicialmente, quando os primeiros trabalhos de visão computacional foram desenvolvidos, os algoritmos e técnicas não conseguiam alcançar resultados muito bons. Os problemas propostos eram muito difíceis de serem resolvidos, isso acontecia tanto pela ausência de técnicas e algoritmos para a finalidade quanto pela limitação dos computadores disponíveis na época. À medida que o tempo passou foram criados *benchmarks* que permitiram avaliar com mais exatidão o estado da visão computacional e comparar os diversos algoritmos existentes.

Em 2012 ocorreu um dos maiores marcos para visão computacional quando a primeira rede neural convolucional, AlexNet, foi apresentada ao mundo. A partir desse momento as redes neurais convolucionais se mantiveram como a melhor solução para problemas de visão computacional. Em 2015 houve outro marco quando a rede neural

convolucional, ResNet, conseguiu resultados melhores que humanos em um desafio de classificação de imagens. Somente em 2020 a técnica de *Vision Transformers* conseguiu resultados melhores que as redes neurais convolucionais em condições adequadas.

Sendo assim, este trabalho avaliou as técnicas que já provaram sua eficiência ao longo do tempo. Ou seja, os modelos de redes neurais, em especial o modelo de rede neural convolucional com arquitetura ResNet. Também avaliamos a performance do *Vision Transformers* pois é a tecnologia mais recente e obtiveram os melhores resultados nos principais *benchmarks*. Foram adotadas 4 abordagens para resolver o problema de marcação de pontos anatómicos em telerradiografias.

1. **Perceptron de Multicamadas:** Uma rede neural de multicamadas é a estrutura mais simples de rede neural. Foi criada uma rede neural de multicamadas e o desempenho da rede foi comparado aos demais modelos de identificação de pontos avaliados. Essa rede recebe como parâmetro a imagem e como resultado deve fornecer os pontos anatômicos da imagem. As imagens utilizadas para treinamento foram normalizadas e centradas em zero.

2. **Rede Neural Convolucional:** As redes neurais convolucionais são extremamente boas para solução de problemas de visão computacional. Dentre elas a ResNet (Residual Network) é uma arquitetura de rede neural profunda que resolve o problema do desvanecimento do gradiente em redes profundas ao introduzir conexões residuais entre as camadas. Essas conexões permitem que a informação flua diretamente do início da rede para o final, sem passar por muitas camadas intermediárias. Isso permite que a rede seja treinada com mais camadas. A partir da arquitetura ResNet criou-se uma rede neural convolucional, que receba como entrada a imagem, a processe em camadas internas e tenha como saída os pontos anatômicos da telerradiografia. Os hiperparâmetros da rede como quantidade de camadas, quantidade de neurónios, tamanho do kernel, etc. foram avaliados durante os experimentos computacionais realizados. As imagens utilizadas para treinamento foram normalizadas e centradas em zero.

3. **Mask R-CNN:** Foi utilizado o framework Mask R-CNN disponibilizado por He em (He et al., 2017). Foram utilizadas as configurações de hiperparâmetros padrões do framework. O framework utiliza uma variação da ResNet em uma arquitetura sofisticada que combina outras redes neurais para propor regiões de interesse na imagem. Essa estratégia de proposta de regiões permite que esse framework obtenha resultados ainda melhores que redes CNNs convencionais. O framework realiza a normalização e a centralização em zero das imagens no treinamento. Esse framework também pode ser usado para outras tarefas de visão computacional como segmentação de imagens, mas nesse trabalho ele será utilizado para identificação de pontos.

4. **Vision Transformer:** Foi implementado um modelo de Vision Transformers para encontrar os pontos anatômicos em uma telerradiografia. Foi utilizado o framework TensorFlow para implementar esse modelo e os melhores hiperparâmetros foram avaliados durante a fase de experimentos. As imagens utilizadas para treinamento foram normalizadas e centradas em zero.

4.3 Decisões técnicas

O conhecimento em computação se materializa na forma de código e muito do que já foi desenvolvido está disponível na forma de bibliotecas e frameworks. A área de visão computacional não é uma exceção e muito do que já foi descoberto está disponível e pode ser utilizado para tornar possível pesquisas subsequentes.

Para criação do banco de dados utilizado nesta pesquisa foi utilizado a linguagem de programação Ruby para criar o extrator de dados da plataforma Cfaz.net. A linguagem Ruby foi utilizada pois o sistema foi implementado nessa linguagem e utilizar a mesma linguagem facilitaria o acesso aos dados. Posteriormente os dados foram organizados na estrutura necessária para o treinamento dos modelos utilizando a linguagem de programação Python com auxílio do Jupyter Notebook. O Python foi a linguagem de programação escolhida por já ser largamente utilizada para ciência de dados e possuir uma grande quantidade de bibliotecas relacionadas.Para analisar e manipular os dados, usou-se a biblioteca Pandas e Numpy.

Especificamente para implementar os modelos de Rede Neural e de Vision Transformer utilizou-se o *framework* TensorFlow. O TensorFlow é um framework *open source* implementado pela Google e é um dos *frameworks* mais utilizados para implementação de redes neurais. Em sua versão 2.0 ele introduziu a interface Keras que simplificou muito a implementação de redes neurais com as mais diversas configurações e hiperparâmetros. Foi escolhido o *framework* TensorFlow porque ele possui um ótimo suporte da comunidade e a sua documentação é bem compreensível.

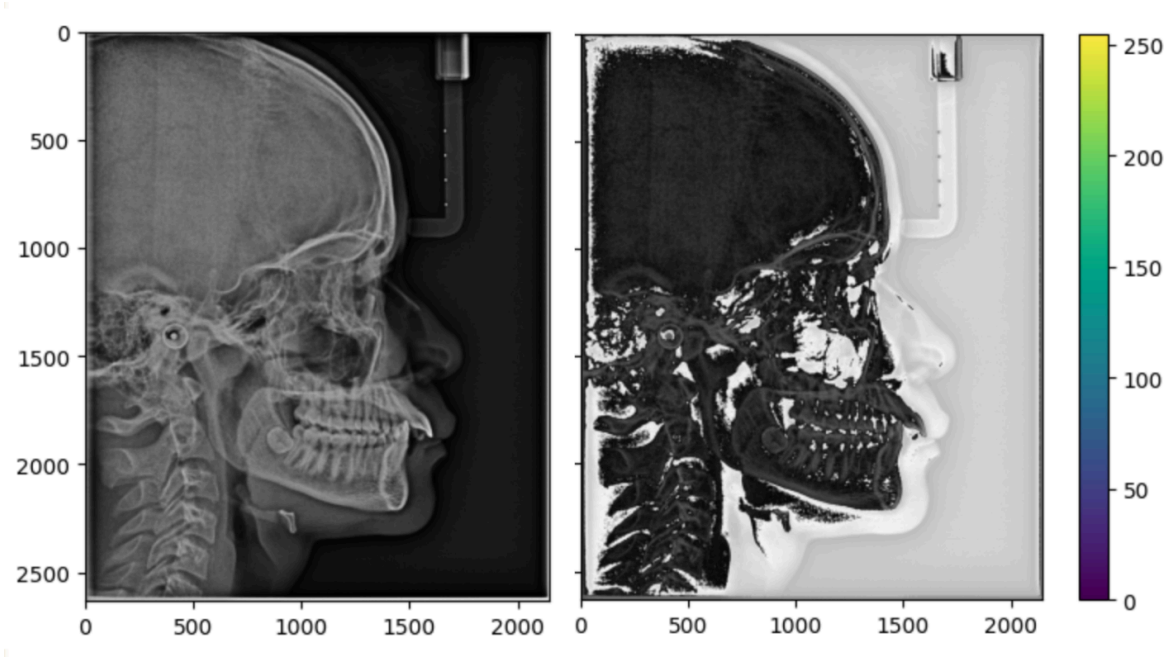
Utilizou-se também o *framework* Mask R-CNN que implementa uma variação da rede neural ResNet utilizando Python e o *framework* PyTorch. O PyTorch é um *framework* open source de aprendizado de máquina implementado pelo Facebook e também é muito utilizado para implementação de redes neurais. O Mask R-CNN foi escolhido porque os pesquisadores disponibilizaram o código e o mesmo pode ser customizado para ser aplicado no problema de detecção de pontos anatômicos em telerradiografias.

Para implementação da variação de Vision Transformer utilizado neste trabalho foi utilizado parte da implementação de vision transformer disponível na biblioteca Transformers disponibilizada pela empresa Hugging Face, essa implementação oferece uma variedade de modelos pré-treinados e configurações de arquitetura, facilitando a experimentação e adaptação para contextos específicos. A escolha dessa biblioteca deve-se à sua ampla adoção na comunidade científica, bem como à sua robustez e eficiência, proporcionando uma base sólida para as modificações e experimentos conduzidos neste estudo.

Para a etapa de pré-processamento dos experimentos, foi adotada uma abordagem que inclui a normalização e a centralização em zero das imagens radiográficas. A normalização consiste na transformação dos valores de intensidade de cada pixel para uma escala entre 0 e 1. Esse processo é essencial para mitigar variações significativas na intensidade e no contraste das imagens, facilitando assim a aprendizagem dos modelos de Inteligência Artificial (IA) ao focar nas características essenciais das imagens. Por sua vez, a centralização em zero é realizada subtraindo a média dos valores dos pixels de toda a imagem, resultando em valores de pixels centralizados em torno de zero. Tal prática é recomendada pela literatura e demonstra ser vantajosa, pois modelos de IA tendem a ter uma melhor performance quando os dados

estão centralizados desta forma. A Figura 32 ilustra o impacto dessas técnicas de pré-processamento em uma telerradiografia da base de dados.

Figura 32 - Pré Processamento da base de dados



Fonte: Autoria própria, 2023

4.4 Traçado cefalométrico para referência comparativa

A maneira mais utilizada para avaliar o desempenho de um modelo é comparar o erro médio desse modelo em um conjunto de testes e mensurar o somatório desse erro. Essa estratégia foi utilizada na maioria dos trabalhos acessados e também é a forma de avaliação utilizada pelos desafios de visão computacional.

Durante a etapa de treinamento, foi avaliado o desempenho dos modelos em relação ao banco de teste de maneira convencional. Não obstante, também será criada uma maneira de comparação dos modelos que seja mais palatável a profissionais da área de radiologia odontológica. Com isso poderemos visualizar todos os modelos aplicados a um mesmo exame e compará-lo diretamente com especialistas humanos. Essa abordagem foi adotada para permitir que os profissionais de radiologia odontológica possam ter uma referência de comparação com a qual estão mais habituados do que as estratégias de avaliação atualmente utilizadas pelos pesquisadores de aprendizado de máquina.

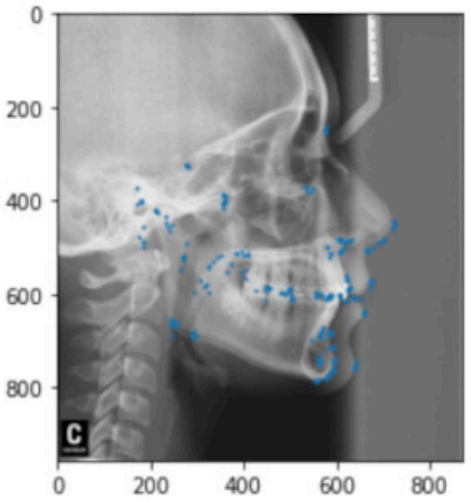
Para realizar a avaliação selecionou-se uma telerradiografia obtida com o aparelho Pax-i da empresa Vatech. A radiografia escolhida é de um paciente adulto com dentição permanente completa e sem nenhum tipo de anomalia óssea.

Foi solicitado a 04 (quatro) especialistas que atuam no mercado de radiologia odontológica para realizar o exame de traçado cefalométrico. Esses exames foram feitos utilizando marcação manual de pontos disponível no software de análise cefalométrica Cfaz.net. Para definir os pontos cefalométricos analisados, utilizamos a literatura da análise da USP. Escolhemos essa análise pois ela é a análise mais utilizada no Brasil, presente em mais de 30% dos exames realizados, conforme o levantamento realizado na Seção 2.3.

A Figura 33 mostra a posição de cada ponto. Nela podemos ver todos os pontos marcados pelos especialistas na forma de tabela juntamente com a representação na imagem.

Figura 33 - Pontos marcados na telerradiografia

	Número	Nome	x	y	Especialista
0	2	N - Násio	577.216	251.155	Especialista 1
1	3	Or - Orbital	534.404	371.860	Especialista 1
2	11	Po - Pório	180.654	397.846	Especialista 1
3	23	Go - Gônio	253.016	672.197	Especialista 1
4	24	Me - Mentoniano	554.951	784.627	Especialista 1
...	...	...	...	...	...
99	35	P' - Ponto P Linha	605.397	482.939	Especialista 4
100	89	V - Ponto V	563.190	751.722	Especialista 4
101	90	T - Ponto T	572.481	680.487	Especialista 4
102	91	Tuber - Tuber	403.168	547.824	Especialista 4
103	92	Pi - Protuberância Incisal	626.682	588.087	Especialista 4

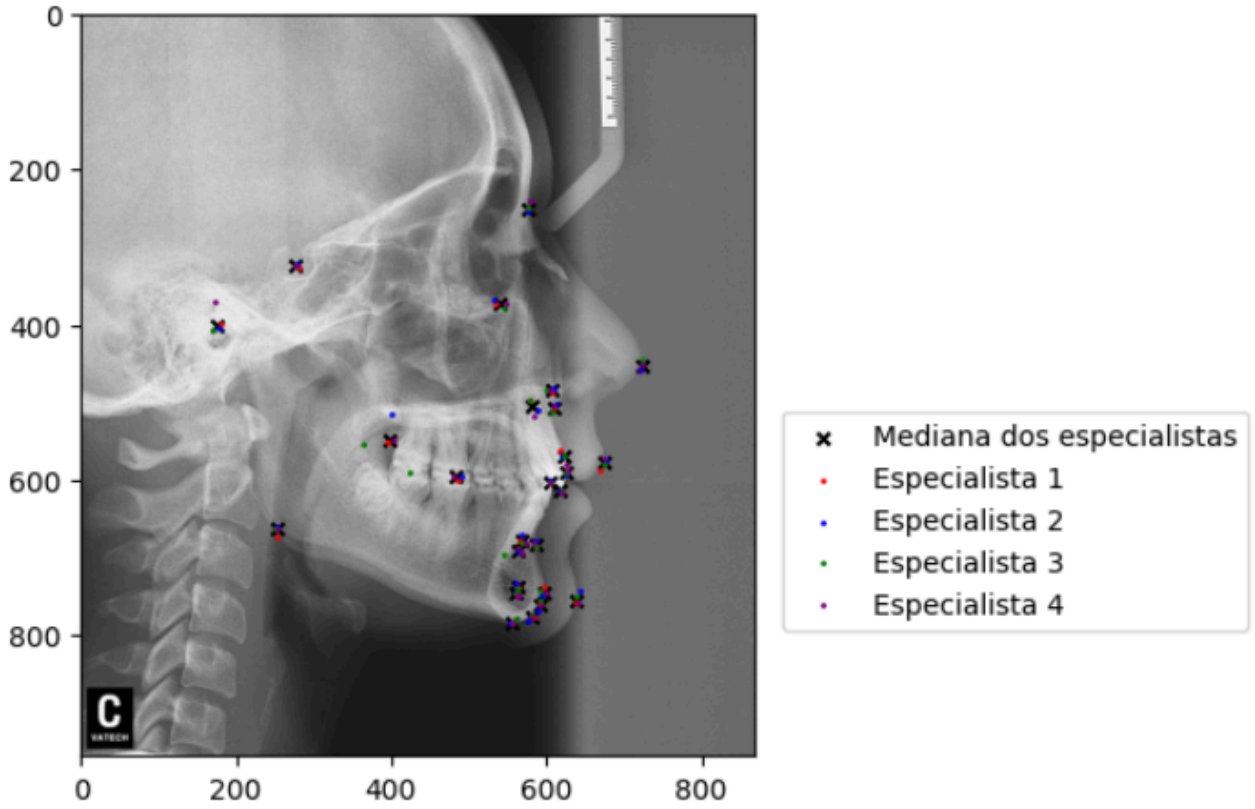


Fonte: Autoria própria, 2022

Não existe uma maneira de definir a posição correta de um ponto cefalométrico, visto que a marcação do ponto depende da interpretação da radiografia pelo especialista. Para determinar a qualidade da marcação de pontos feita por cada especialista foi criada uma referência própria, obtida através da mediana obtida com os pontos marcados por todos os especialistas. A Figura 34 apresenta as anotações cefalométricas feitas por

diferentes especialistas, com cada um representado por uma cor distinta, conforme indicado na legenda. A mediana de referência adotada para comparação é destacada pelo símbolo 'X'.

Figura 34 - Referência adotada para comparação



Fonte: Autoria própria, 2023

Essa referência foi usada para determinar o erro de cada ponto marcado pelos especialistas. O erro foi calculado pela distância euclidiana em milímetros do ponto marcado até o ponto de referência. Posteriormente vamos utilizar o erro de cada um dos modelos avaliados para compará-los a cada especialista.

Na Figura 35 abaixo podemos ver cada ponto com seu respectivo erro em milímetros.

Figura 35 - Erro de cada ponto cefalométrico

	Número	Nome	x	y	Especialista	Erro em pixels	Erro em milímetros
0	2	N - Násio	577.216	251.155	Especialista 1	1.282349	0.290636
1	3	Or - Orbital	534.404	371.860	Especialista 1	4.808457	1.089804
2	11	Po - Pório	180.654	397.846	Especialista 1	6.721966	1.523488
3	23	Go - Gônio	253.016	672.197	Especialista 1	9.770424	2.214401
4	24	Me - Mentoniano	554.951	784.627	Especialista 1	0.760581	0.172381
...	...	...	...	...	...	...	...
99	35	P' - Ponto P Linha	605.397	482.939	Especialista 4	1.329087	0.301229
100	89	V - Ponto V	563.190	751.722	Especialista 4	6.435793	1.458629
101	90	T - Ponto T	572.481	680.487	Especialista 4	4.415579	1.000761
102	91	Tuber - Tuber	403.168	547.824	Especialista 4	5.485976	1.243360
103	92	Pi - Protuberância Incisal	626.682	588.087	Especialista 4	20.125701	4.561355

Fonte: Autoria própria, 2022

Com o erro individual de cada ponto, podemos então somá-lo aos erros obtidos em cada ponto, de modo que esse somatório representa o erro total da análise cefalométrica. Fazendo isso, podemos ver a dimensão do erro total cometido por cada especialista e pela inteligência artificial, o que nos leva a um número que permite fazer a comparação de resultados.

5 Resultados

Neste capítulo, apresentam-se os resultados das diferentes técnicas de visão computacional consideradas mais adequadas, aplicadas à tarefa de identificação de pontos anatômicos. Essas técnicas, bem como a motivação para a escolha de cada uma delas, foram descritas na Seção 4.2.

Para cada uma dos modelos criados serão discutidas as escolhas de implementação e será apresentado o erro obtido no conjunto de treinamento e de validação. O modelo criado também será utilizado para fazer o traçado cefalométrico de referência comparativa descrito na Seção 4.4 para que a capacidade do modelo possa ser comparada a especialistas reais.

5.1 Perceptron de Multicamadas

A rede neural de multicamadas que apresentou melhor resultado na tarefa de encontrar pontos anatômicos em telerradiografias foi implementada com 3 camadas ocultas. Cada camada contendo 128 neurônios completamente conectados e com função de ativação *Relu*. A cada uma das camadas ocultas foi adicionado um *Dropout* de 0,1. Como dados de entrada da rede neural, foram utilizadas imagens com 64 pixels de largura e altura, elas foram convertidas em um vetor linear de pixels que são processadas pela rede. A última camada da rede é formada por uma camada com 52 neurônios, um para cada coordenada x e y dos 26 pontos de interesse. Nessa camada foi utilizada a função de ativação sigmóide. A avaliação da precisão do modelo foi realizada por meio do cálculo do erro médio quadrático e a metodologia de descida do gradiente adotada foi o otimizador Adam.

Às imagens do conjunto de dados de treinamento apresentado no Capítulo 4.1 foram aplicadas as técnicas de pré processamento de centralização em zero e normalização. A normalização consiste em transformar os dados de forma que eles tenham uma escala comum. Isso é feito para evitar que uma característica com valores muito grandes tenha um peso muito maior que outras características na análise dos dados. Já o zero-centering é uma técnica que transforma os dados para que a média deles seja zero. Isso é feito para que o modelo de aprendizado de máquina possa

aprender mais facilmente padrões em relação aos dados. Também foi realizada a normalização dos valores das coordenadas dos pontos de interesse.

O conjunto de dados disponível foi separado em conjunto de treinamento e conjunto de teste, foram separadas as últimas 1000 imagens para o conjunto de treinamento. No processo de treinamento da rede neural, o conjunto de dados de treinamento foi dividido em conjunto de treinamento e conjunto de validação com proporção de 9 para 1. Ao final do treinamento, que foi realizado nas 29 mil imagens do conjunto de treinamento, obteve-se um erro médio quadrático de 3.0921e-04. Já no conjunto de dados de teste, o erro médio quadrático foi de 3.7395e-04. Esses resultados indicam que a rede neural não sofreu de *overfitting*, uma vez que as taxas de erro obtidas no grupo de validação e de treinamento foram similares.Também pode-se constatar que o tamanho da base de dados foi muito superior ao necessário para treinar a rede, uma vez que os valores de erro no conjunto de teste e de validação alcançaram os valores de erro mais baixos com apenas 2 mil imagens.

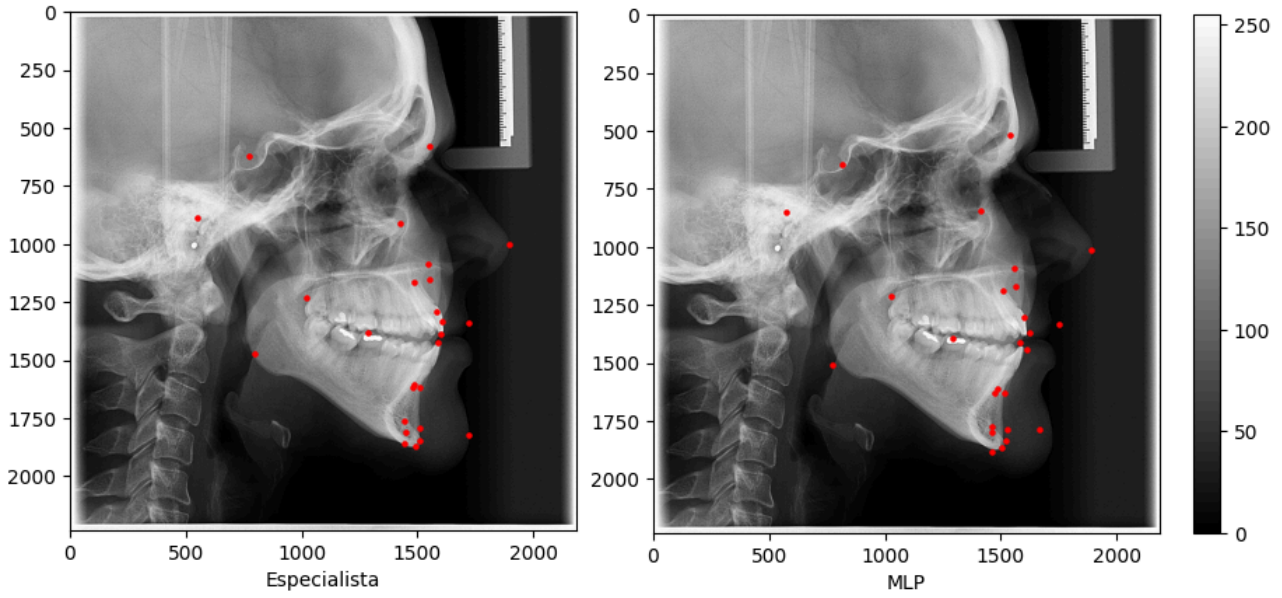
Durante o processo de teste da rede neural, foram realizados experimentos variando o número de camadas, o número de neurônios em cada camada e o tamanho da imagem de entrada. Foi observado que um maior número de neurônios não resultou em uma melhoria significativa no erro, enquanto um número menor de neurônios fez com que a rede não convergisse para um valor de erro semelhante. A diminuição ou aumento do número de camadas não melhorou o erro da rede. Além disso, o aumento do tamanho da imagem de entrada para 128px ou 256px não melhorou os resultados, mesmo com o aumento do número de neurônios. Uma possível explicação para isso é que a redução da imagem realizou uma operação semelhante ao *max pooling*, incorporando mais informações sobre a imagem em cada pixel, uma técnica comumente usada em redes neurais convolucionais. Não foi possível testar tamanhos de imagem maiores, pois não havia recursos de hardware suficientes para realização dos testes.

Sobre a função de ativação da camada de saída, observou-se que a função *Sigmoid* obteve resultados muito melhores do que a Relu ou Softmax. Uma possível explicação para esse resultado é que os valores de saída da rede são constantes e foram normalizados no pré-processamento do conjunto de treinamento, sendo assim, faz sentido que uma função que produza valores entre 0 e 1 obtenha bons resultados.

O *Dropout* aplicado às camadas ocultas não mostrou nenhum impacto significativo nos resultados, isso ocorreu pois o tamanho da base de dados era suficientemente grande para garantir uma boa generalização dos dados e para evitar o ajuste demasiadamente aos dados de treinamento. Entretanto, o *Dropout* foi mantido na arquitetura da rede, pois essa técnica de regularização é amplamente utilizada em redes neurais para reduzir o overfitting.

A Figura 36 mostra os pontos do conjunto de treinamento marcados por um especialista à esquerda e os pontos preditos pela rede MLP à direita.

Figura 36 - Comparativo pontos do conjunto de dados e pontos previstos pela MLP



Fonte: Autoria própria, 2023

5.2 Rede Neural Convolucional

A arquitetura de rede convolucional escolhida para tarefa de identificação de pontos anatômicos nas telerradiografias foi a ResNet (*Residual Neural Network*). Essa arquitetura de rede foi responsável por uma grande resolução nas tarefas de visão computacional e quando foram introduzidas foram consideradas o estado da arte para tarefas de visão computacional. A ResNet é composta por várias camadas de convolução, *batch normalization* e ativação ReLU. A arquitetura da ResNet é baseada em blocos residuais, que permitem que as camadas aprendam funções residuais em vez de tentar aprender a mapear a entrada diretamente para a saída. Cada bloco residual tem duas

camadas convolucionais com um atalho de conexão que pula uma ou mais camadas para fornecer uma via de informação alternativa. Foi escolhida a ResNet50 que contém 50 camadas e usa blocos residuais com duas camadas convolucionais com 64, 128, 256 e 512 filtros, respectivamente. A camada de entrada da ResNet é uma camada convolucional que processa a imagem de entrada com o número correto de canais de cor (geralmente 3). A camada de saída é uma camada densa de 1000 neurônios que produz a saída da rede com o número correto de unidades, dependendo do problema em questão.

A Rede ResNet não utiliza *Dropout* pois a técnica de aprendizado residual já reduz o problema de *overfitting*, além disso, a rede também não usa *max pooling* pois em sua arquitetura profunda o *max pooling* não adiciona nenhuma melhoria na acurácia da rede.

A ResNet50 padrão foi alterada para que como entrada sejam passadas imagens em escala de cinza, ou seja, imagens que tem apenas um canal de cor em vez dos 3 canais de uma imagem rgb. Também alteramos a camada de saída da rede ,que originalmente foi criada para classificação, adicionando uma camada final completamente conectada com 52 neurônios para representar as coordenadas dos 26 pontos de interesse.

Antes do treinamento, os dados do conjunto de treinamento descritos no capítulo 4.1 também foram pré-processados e sobre eles foram aplicadas normalização e centralização em zero.

O conjunto de dados disponível foi separado em conjunto de treinamento e conjunto de teste, foram separadas as últimas 1000 imagens para o conjunto de treinamento. No processo de treinamento da rede neural, o conjunto de dados de treinamento foi dividido em conjunto de treinamento e conjunto de validação com proporção de 9 para 1. Ao final do treinamento, que foi realizado nas 29 mil imagens do conjunto de treinamento, obteve-se um erro médio quadrático de 9.2569e-05. Já no conjunto de dados de teste, o erro médio quadrático foi de 9.8464e-05. Esses resultados indicam que a rede neural não sofreu de *overfitting*, uma vez que as taxas de erro obtidas no grupo de validação e de treinamento foram similares. Também pode-se constatar que o tamanho da base de dados foi muito superior ao necessário para treinar a rede, uma vez que os valores de erro no conjunto de teste e de validação alcançaram os valores de erro mais baixos com apenas 6 mil imagens.

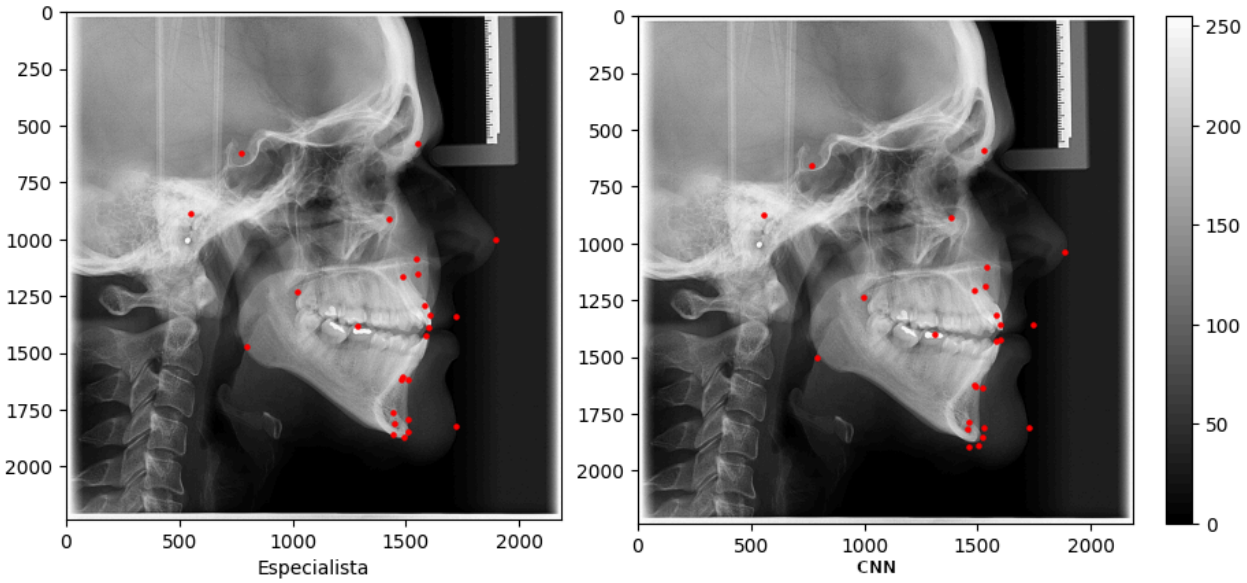
Durante o processo de treinamento foram realizados experimentos variando a entrada de dados para a ResNet. Essa rede foi inicialmente concebida para processar imagens com 3 camadas rgb, então foi realizada uma bateria de treinamento passando lendo as radiografias em formato rgb com 3 camadas, entretanto esse canais adicionais resultaram em melhora na acurácia da rede, por isso os testes finais foram realizados utilizando apenas um canal de cor. Utilizando 3 canais de cor foi possível testar a transferência de aprendizados dos pesos treinados no banco de dados ImageNet, mas esses pesos apesar de acelerarem o processo de treinamento não refletiram em valores menores da função de erro.

Com o objetivo de avaliar a influência da variação na resolução da imagem nos resultados, foram realizados testes com imagens de 64, 128, 256 e 512 pixels. Os resultados indicaram que as resoluções de 128 e 256 pixels apresentaram conversão similar e melhores resultados, enquanto o tamanho de 512 pixels não foi viável devido às limitações de hardware do computador utilizado. Com base nesses resultados, optou-se por utilizar a resolução de 256 pixels para os testes finais, uma vez que a rede ResNet foi inicialmente desenvolvida para processar imagens do banco de dados ImageNet, que possui imagens com resolução de 224x224 pixels.

Além disso experimentou-se adicionar camadas completamente conectadas ao final da rede ResNet, essa é uma prática comumente aplicada quando utilizamos uma rede projetada para outro propósito. Essa estratégia não obteve nenhuma melhora, isso aconteceu porque a rede ResNet já é muito profunda e a adição de mais camadas não trouxe qualquer melhoria à rede.

A Figura 37 mostra os pontos do conjunto de treinamento marcados por um especialista à esquerda e os pontos preditos pela rede CNN à direita.

Figura 37 - Comparativo pontos do conjunto de dados e pontos previstos pela CNN



Fonte: Autoria própria, 2023

5.3 Mask R-CNN

As R-CNN (Redes Neurais Convolucionais com Proposta de Região de Interesse) se mostraram muito eficientes para diversas tarefas de visão computacional sendo amplamente utilizadas para a detecção e identificação de pontos-chave em imagens. Essa arquitetura combina as características poderosas das redes convolucionais com a capacidade de localização precisa de regiões de interesse. Essa estratégia adota uma abordagem em duas etapas para identificar pontos chave em uma imagem.

Primeiro, o modelo gera uma lista de regiões de interesse (RoIs) potenciais na imagem usando uma rede convolucional, que é pré-treinada em um conjunto de dados anotados. Essas regiões de interesse são propostas com base na probabilidade de conterem objetos de interesse. Em seguida, cada RoI é alimentada em uma rede neural especializada, chamada de rede de extração de características regionais, que é treinada para identificar e localizar os pontos-chave dentro dessa região específica. A rede de extração de características regionais utiliza convoluções espaciais e operações de pooling para capturar as informações discriminativas relevantes dentro de cada região de interesse, permitindo a identificação precisa dos pontos-chave. Ao combinar essas duas etapas, o modelo R-CNN é capaz de localizar e identificar com precisão os pontos-chave

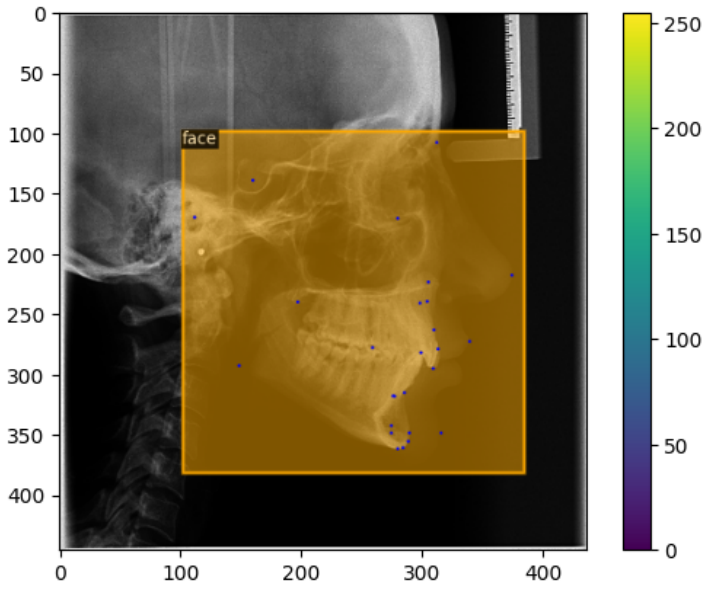
em uma imagem. O funcionamento da arquitetura de rede R-CNN está descrita com mais detalhes no capítulo 3.2.

Para implementar o modelo foi selecionado o *framework* Detectron2. O *framework* é mantido pela equipe do Facebook AI Research (FAIR) e pela comunidade de código aberto. Isso significa que ele recebe atualizações regulares, correções de bugs e contribuições da comunidade, garantindo uma base sólida para realizar experimentos e obter resultados confiáveis. A escolha do Detectron2 também foi influenciada pela sua ampla adoção e pela comunidade ativa que o suporta. Sua implementação é eficiente e extensível facilitando a adaptação para a identificação de pontos-chave em telerradiografias.

O modelo utilizado foi o R101-FPN disponibilizado pela equipe do Facebook AI Research. Esse modelo foi pré-treinado em um grande conjunto de dados de referência, o COCO: Common Objects in Context, que contém imagens anotadas com diversos pontos-chave em diferentes contextos. Esse modelo obteve o melhor resultado para identificação de pontos chaves na base de dados COCO dentre os modelos de pontos chave disponíveis. Assim, o uso desse modelo pré-treinado no Detectron2 proporciona uma inicialização sólida para a identificação de pontos-chave em telerradiografias.

A base de dados de telerradiografias e pontos chaves foi organizada para atender a estrutura estabelecida pelo framework. Para cada imagem foi necessário criar um dicionário de dados com python contendo o caminho para a imagem, uma caixa de seleção que engloba o objeto para treinar a primeira etapa do modelo e os pontos anotados para o treinamento da segunda etapa do modelo. O objeto escolhido para identificação foi a face que contém todos os pontos. A figura abaixo mostra um exemplo da base de dados organizada.

Figura 38 - Dicionário de dados para R-CNN



Fonte: Autoria própria, 2023

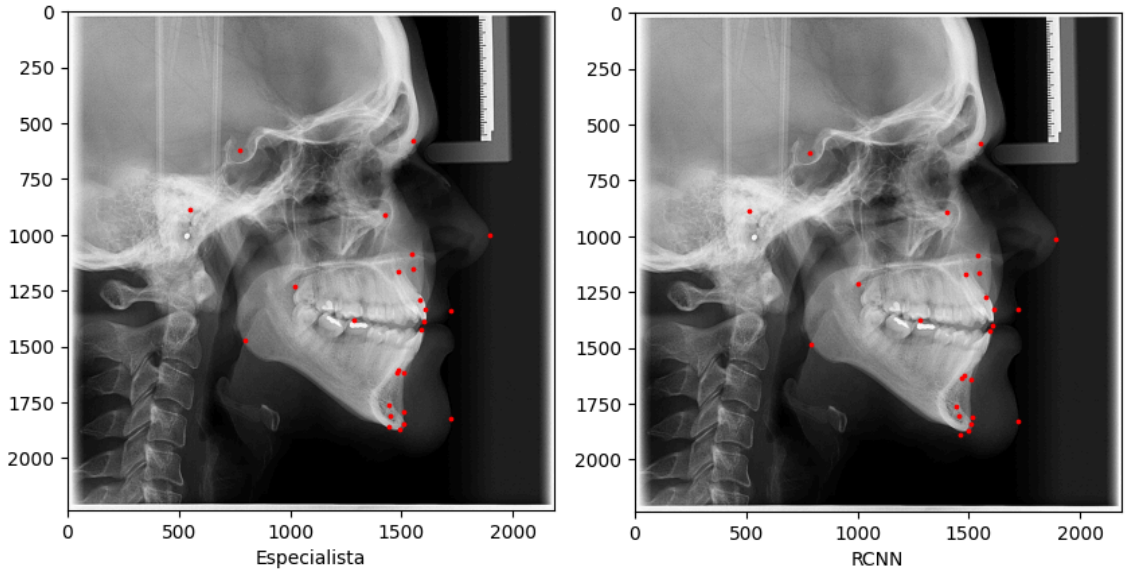
O conjunto de dados disponível foi separado em conjunto de treinamento e conjunto de teste, foram separadas as últimas 1000 imagens para o conjunto de treinamento. Ao final do treinamento, que foi realizado nas 29 mil imagens do conjunto de treinamento, obteve-se um erro médio quadrático de 1.2968. Já no conjunto de dados de teste, o erro médio quadrático foi de 1.4171. Esses resultados indicam que a rede neural não sofreu de *overfitting*, uma vez que as taxas de erro obtidas no grupo de validação e de treinamento foram similares. Também pode-se constatar que o tamanho da base de dados foi muito superior ao necessário para treinar a rede, uma vez que os valores de erro no conjunto de teste e de validação alcançaram os valores de erro similares com apenas 10 mil imagens sem apresentar *overfit*.

Durante o processo de treinamento foram usados hiperparâmetros similares aos valores de treinamento do framework Detectron2. Esses valores foram escolhidos porque já se provaram os mais eficientes na detecção de pontos no conjunto de dados COCO. O único parâmetro que foi variado foi o tamanho máximo da imagem de entrada, que no modelo original possuía 1333px, mas no treinamento deste trabalho foi reduzido para 800px devido a limitações de hardware para treinamento. Os testes de variação de tamanho, entretanto, não apontam que um tamanho maior na imagem de entrada traga

resultados melhores, foi usado 800px por ser o mais próximo possível do utilizado no treinamento padrão do modelo.

A Figura 39 mostra os pontos do conjunto de treinamento marcados por um especialista à esquerda e os pontos preditos pela rede CNN à direita.

Figura 39 - Comparativo pontos do conjunto de dados e pontos previstos pela R-CNN



Fonte: Autoria própria, 2023

5.4 Vision Transformer

A aplicação de transformers para tarefas de visão computacional são recentes e prometem revolucionar a visão computacional da mesma maneira que revolucionaram o processamento de linguagem natural. Neste capítulo serão apresentados os testes da aplicação dessa tecnologia para encontrar pontos anatômicos em telerradiografias.

Na busca por soluções de vanguarda para reconhecimento de imagens, foi adotada a implementação proposta e validada no artigo "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" (DOSOVITSKIY, 2020). Esta pesquisa, por sua vez, introduziu uma inovação notável ao empregar transformers para processamento de imagens, demonstrando performance que superou as arquiteturas de redes neurais convolucionais (CNNs) tradicionais no dataset ImageNet. A escolha pelos parâmetros especificados no referido estudo justifica-se pela extensa avaliação e otimização realizada

pelos autores, que investigaram variações na arquitetura e chegaram a resultados de excelência. Esta premissa elimina a necessidade de ajustes adicionais, uma vez que os parâmetros selecionados já demonstraram ser eficientes em contextos similares. Além disso, realizar otimização dos hiperparâmetros da arquitetura de transformer apresentada estão fora do escopo do projeto.

No referido artigo, os pesquisadores adotaram uma arquitetura de modelo caracterizada por 12 camadas de transformadores, cada uma dotada de 12 cabeças de atenção e um tamanho de embedding oculto estabelecido em 768. A dimensão intermediária da rede *feed-forward* dentro de cada bloco do transformador é 3072. Além disso, os *patches* das imagens, que representam segmentações para análise, são definidos com dimensões de 16x16 pixels. Este formato foi criteriosamente selecionado para capturar detalhes suficientes enquanto mantém a eficiência computacional. Para um entendimento mais aprofundado sobre a mecânica e os princípios subjacentes à arquitetura dos transformers, recomenda-se a consulta à Seção 2.4.4, onde é apresentada uma fundamentação teórica sobre o tema.

A adaptação do modelo proposto no artigo ao desafio em questão envolveu a integração de uma camada densa à saída do modelo base apresentado no artigo. Esta camada visa mapear as representações de alto nível em coordenadas específicas na imagem. A camada densa incorporada contém N neurônios, com N representando as coordenadas dos pontos a serem identificados. Utilizou-se uma função de ativação sigmoidal para delimitar a saída a um intervalo específico. A função de ativação sigmoidal foi escolhida para realizar a classificação binária, produzindo valores no intervalo entre 0 e 1 que representam as coordenadas dos pontos. Esta modificação garante que as saídas do modelo se alinhem diretamente às coordenadas dos pontos de interesse presentes nas radiografias.

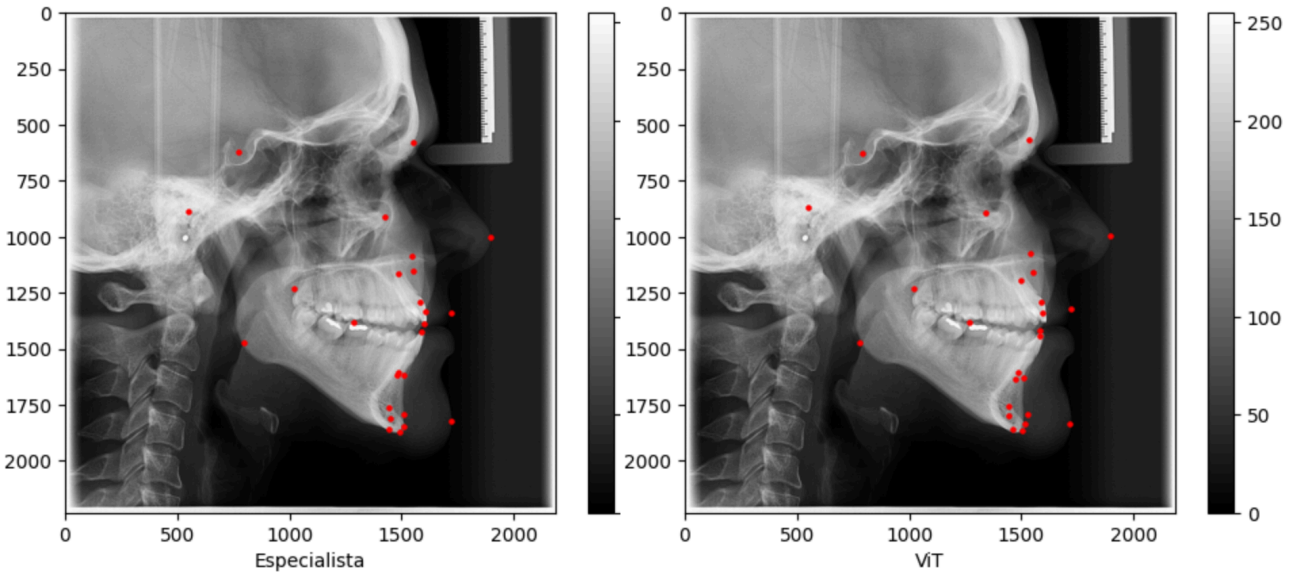
No processo de treinamento, optou-se pela utilização de pesos previamente treinados na ImageNet, que contempla um conjunto de dados composto por aproximadamente 21 milhões de imagens. A motivação para tal escolha residia na complexidade e magnitude do modelo em questão, cujo tempo de treinamento se revela substancialmente elevado, especialmente quando contrastado com outros modelos avaliados no escopo deste estudo. As imagens empregadas no treinamento foram configuradas com dimensões de 224px tanto em altura quanto em largura, além de serem

compostas por 3 canais de cores. Tal especificação decorre das características intrínsecas do modelo base adaptado para este estudo. Ressalta-se que não estava no escopo desta pesquisa realizar avaliações ou propostas de otimização na arquitetura do transformer apresentado no artigo de referência.

O conjunto de dados disponível foi separado em conjunto de treinamento e conjunto de teste, foram separadas as últimas 1000 imagens para o conjunto de treinamento. No processo de treinamento da rede neural, o conjunto de dados de treinamento foi dividido em conjunto de treinamento e conjunto de validação com proporção de 9 para 1. Ao final do treinamento, que foi realizado nas 29 mil imagens do conjunto de treinamento, obteve-se um erro médio quadrático de 2.0568e-05. Já no conjunto de dados de teste, o erro médio quadrático foi de 6.0005e-05. Esses resultados indicam que a rede neural não sofreu de *overfitting*, uma vez que as taxas de erro obtidas no grupo de validação e de treinamento foram similares. Também pode-se constatar que o tamanho da base de dados foi muito superior ao necessário para treinar a rede, uma vez que os valores de erro no conjunto de teste e de validação alcançaram os valores de erro mais baixos com apenas 10 mil imagens.

A figura 40 mostra os pontos marcados por um especialista à esquerda ao lados dos pontos preditos pela rede ViT à direita:

Figura 40 - Comparativo pontos do conjunto de dados e pontos previstos pela rede ViT



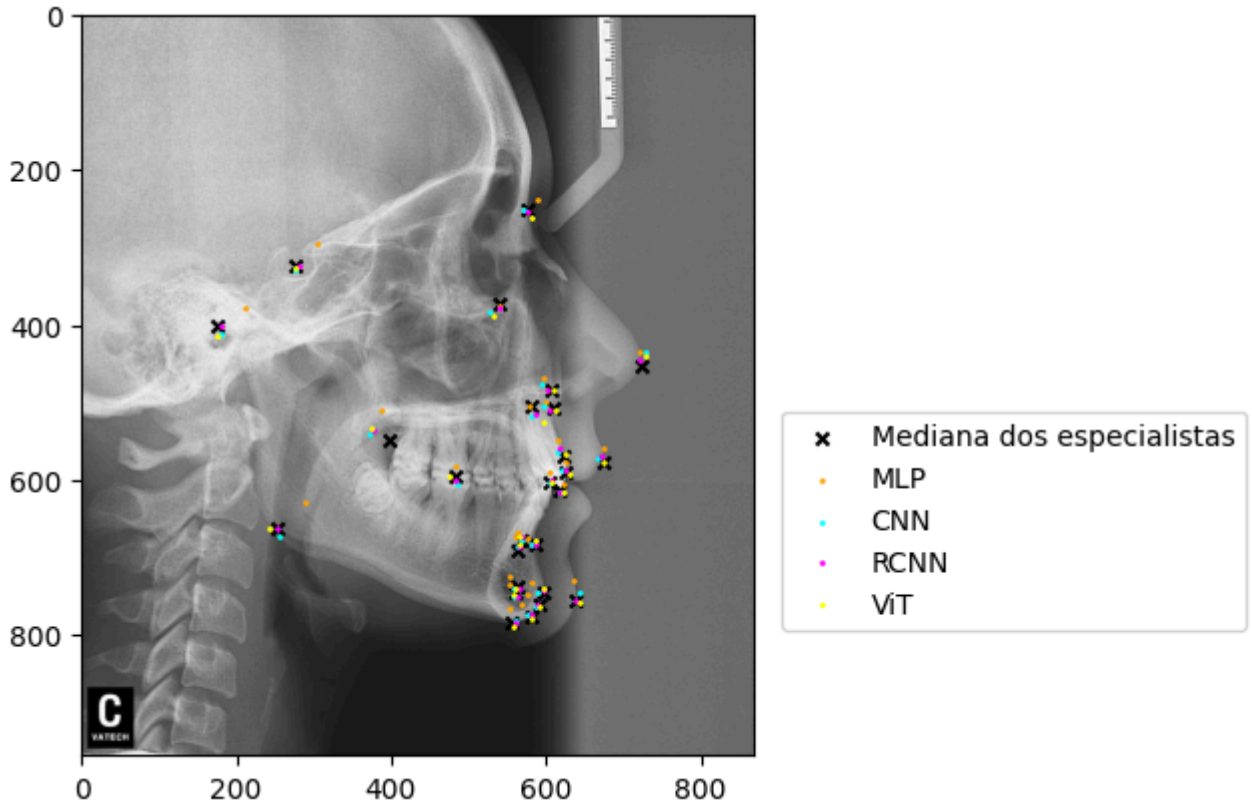
Fonte: Autoria própria, 2023

5.5 Traçado cefalométrico para referência comparativa

O racional por trás da construção do traçado cefalométrico para referência comparativa foi apresentado no capítulo 4.4.

A Figura 41 exibe as anotações cefalométricas realizadas por todos os modelos treinados, sendo que cada modelo é representado por uma cor específica, conforme detalhado na legenda. O símbolo 'X' destaca a mediana de referência adotada para comparação.

Figura 41 - Pontos anotados na cefalometria



Fonte: Autoria própria, 2023

A Figura 42 mostra o erro acumulado dos pontos marcados por especialista comparado ao erro acumulado obtido com os pontos identificados pelos modelos de visão computacional criados neste trabalho.

Figura 42 - Erro acumulado por especialista

Especialista	Somatório do erro em milímetros
Especialista 1	22.613103
RCNN	32.262145
Especialista 2	34.680282
Especialista 4	35.406523
Especialista 3	44.495072
ViT	45.176033
CNN	51.552723
MLP	114.498645

Fonte: Autoria própria, 2023

Conforme apresentado na Figura 41, pode-se observar os erros acumulados de diversos modelos e especialistas em comparação à referência estabelecida. Nota-se que a rede neural CNN registrou um erro acumulado de 51.552723, o qual, ainda que seja inferior ao da rede MLP, com erro de 114.498645, demonstrou ser superior ao dos especialistas. Observa-se também que o modelo R-CNN exibiu um erro acumulado de 32.262145, posicionando-se favoravelmente em relação aos modelos CNN, MLP e ViT. Além disso, este modelo superou três dos especialistas em termos de precisão, especificamente Especialista 2, Especialista 3 e Especialista 4, com erros acumulados de 34.680282, 44.495072 e 35.406523, respectivamente. No que concerne ao modelo ViT, com um erro de 45.176033, este demonstrou um desempenho superior tanto ao MLP quanto à CNN. Entretanto, o ViT não conseguiu alcançar a eficiência do modelo R-CNN e tampouco superar os erros acumulados registrados pelos especialistas.

5.6 Discussão

Os resultados indicam que os modelos de visão computacional apresentaram desempenho similar ao dos especialistas avaliados. As representações gráficas das previsões dos modelos propostos corroboram essa constatação, demonstrando a

proximidade dos valores estimados aos valores identificados por especialistas. Ou seja, o traçado automático se mostrou uma ferramenta valiosa para auxiliar a realização de exames de análise cefalométrica.

Os algoritmos MLP, CNN e ViT exibiram precisão inferior em comparação à referência estabelecida pelos especialistas descrita no capítulo 5.5. No entanto, o modelo R-CNN superou o desempenho de alguns especialistas pela metodologia de comparação utilizada. Surpreendentemente, o modelo ViT, apesar de ser uma tecnologia emergente em classificação de imagem, não liderou os resultados na presente pesquisa.

A precisão inferior do modelo MLP era antecipada, dado que redes neurais multicamadas não são primariamente designadas para reconhecimento de imagem. Este estudo reitera tal limitação, evidenciando sua inadequação para detecção precisa de pontos anatômicos em telerradiografias.

A precisão do modelo CNN superou o MLP, um resultado previsível considerando que as CNNs são especificamente projetadas para tarefas de visão computacional e são reconhecidas por sua alta eficácia neste domínio.

A precisão superior do modelo Vision Transformer (ViT) em relação à CNN é notável. Isso porque o funcionamento do ViT segmenta a imagem em patches de 16px e subsequente os converte em tokens. A estes tokens são associadas informações de posicionamento, permitindo que o processamento pelo transformador identifique relações entre tokens, independentemente de suas distâncias iniciais. Contrapondo, a rede CNN inicia sua operação através da convolução de pixels com um kernel, focalizando inicialmente nas informações dos pixels adjacentes e somente em camadas subsequentes analisa relações mais distantes na imagem. Enquanto o mecanismo do ViT é eficaz em tarefas de classificação devido à sua capacidade de identificar um objeto com base em diferentes partes da imagem, a conversão de patches em tokens sugeria potencial perda de detalhes minuciosos, como a posição precisa de um ponto. Contudo, a superioridade do ViT em relação à CNN pode ser justificada pelo fato de que a localização de um ponto anatômico em uma cefalometria lateral da cabeça é influenciada não apenas por estruturas imediatamente adjacentes ao ponto, mas pelo conjunto total das estruturas cranianas.

A precisão superior do modelo R-CNN pode ser atribuída à sua complexidade arquitetônica, que integra múltiplas redes neurais para identificação dos pontos.

Inicialmente, ele identifica a região de interesse, restringindo assim a área de busca, aumentando a precisão. Adicionalmente, a R-CNN emprega uma CNN para prever uma máscara de segmentação, que tem como objetivo identificar os pixels do ponto anatômico desejado, a máscara prevista contendo a probabilidade de cada ponto ser o ponto desejado cria uma espécie de mapa de calor que identifica a posição do ponto. Esta metodologia refinada do R-CNN justifica sua precisão superior em comparação com o modelo ViT, apesar do último apresentar melhor desempenho que a CNN em análises isoladas.

É importante observar a natureza subjetiva do exame de traçado cefalométrico, pois depende da capacidade de localização do ponto de cada especialista, sendo assim torna-se difícil comparar um especialista com outro e afirmar qual é melhor. Dois especialistas que realizarem o mesmo exame irão obter resultados ligeiramente diferentes e ambos estão corretos. Essas variações são esperadas e não trazem prejuízo ao exame, pois não alteram o diagnóstico do dentista que avalia o exame. Sendo assim, o modelo de marcação de pontos automático R-CNN pode ser considerado adequado para realização do exame, uma vez que obteve resultados comparáveis aos de especialistas. O modelo ViT, por ter se aproximado ao resultado dos especialistas, também representa uma possibilidade viável, visto que pequenas variações na marcação não trazem prejuízo ao diagnóstico.

É importante salientar que a base de dados utilizada neste estudo foi coletada a partir de um sistema real. Não foi uma prioridade assegurar a representatividade, no conjunto de treinamento, de condições específicas como a ausência de dentes, anomalias ósseas e diferentes fases do desenvolvimento ósseo, incluindo a dentição mista ou permanente. Consequentemente, essas variáveis devem ser consideradas ao avaliar a capacidade de generalização dos modelos propostos para diversos cenários.

A comprovação de que a inteligência artificial é capaz de obter resultado comparável ao de especialistas para o recorte levantado é relevante, pois este representa a maioria dos exames odontológicos realizados pelas clínicas de radiologia no Brasil. Os resultados apresentados nesta dissertação fornecem uma base sólida para a utilização desses modelos em aplicações clínicas.

6 Considerações Finais

Foi realizado um estudo extensivo sobre o exame de traçado cefalométrico, qual a sua finalidade e quais parâmetros são relevantes para o diagnóstico. Com esse estudo pôde-se entender como um exame de traçado cefalométrico é realizado, em especial utilizando a análise cefalométrica no padrão da USP, por ser a mais utilizada no Brasil foi escolhida como material de estudo neste trabalho.

Também foi feita uma extensiva pesquisa exploratória sobre visão computacional. Foram avaliadas os principais algoritmos de visão computacional, desde os antigos e já obsoletos até os algoritmos que compreendem o estado da arte da visão computacional. Fazendo esse levantamento ficou claro que muitas das técnicas de visão computacional poderiam ser aplicadas para auxiliar a execução de exames de traçado cefalométrico. Além disso, definiu-se estratégias com potencial para realizar um exame de traçado cefalométrico de maneira automática.

O principal objetivo foi aplicar técnicas avançadas de visão computacional, particularmente baseadas em redes neurais, para aprimorar sistemas de auxílio ao diagnóstico, com foco no exame de traçado cefalométrico. Ao analisar o estado da arte, foi possível selecionar as técnicas mais eficazes para a identificação de pontos cefalométricos em telerradiografias laterais. Construimos também uma base de conhecimento robusta a partir de exames realizados por especialistas. Nossos estudos revelaram técnicas de visão computacional que demonstraram desempenho notável em comparação com especialistas humanos no exame em questão. Consequentemente, confirmamos que a utilização de técnicas de visão computacional podem auxiliar e, em algumas circunstâncias, superar a precisão de especialista na elaboração do exame de traçado cefalométrico. O resultado prático desta pesquisa sugere uma possível redução no tempo de elaboração desses exames, garantindo precisão e redução de custos para clínicas, culminando na melhoria da eficiência diagnóstica para o benefício dos pacientes. Esta dissertação, portanto, serve como um marco no emprego de tecnologias emergentes na melhoria da saúde e qualidade de vida da população, mostrando como abordagens interdisciplinares de um problema são viáveis e podem ser benéficas para ambas as áreas.

Nas considerações finais do presente estudo, destaca-se a subjetividade associada ao exame de traçado cefalométrico, em virtude da variabilidade na localização de pontos por especialistas distintos. Esta variabilidade, no entanto, não compromete a validade do exame, uma vez que os diagnósticos clínicos se mantêm consistentes. Quatro modelos de redes neurais foram desenvolvidos, baseando-se nas tecnologias identificadas como promissoras na revisão bibliográfica: um modelo com rede perceptron multicamadas, um com rede neural convolucional, um com rede neural convolucional regional (R-CNN) e um com vision transformers (ViT). Cada modelo foi empregado na identificação dos pontos anatômicos necessários para o exame de traçado cefalométrico conforme o padrão da USP. A precisão de cada modelo foi avaliada comparando os pontos identificados pelos modelos com aqueles marcados por especialistas, utilizando a distância entre o ponto marcado pelo modelo e a mediana dos pontos dos especialistas como métrica de erro. Essa comparação permitiu uma avaliação comparativa da eficácia dos modelos.

Os modelos R-CNN e ViT mostraram-se eficazes, apresentando desempenhos similares aos de especialistas.. A aptidão da inteligência artificial para produzir resultados comparáveis aos de especialistas, dentro do recorte proposto, revela-se relevante, dado que tal perfil representa a maioria dos exames odontológicos realizados no Brasil. Esta pesquisa, portanto, estabelece um precedente robusto para a integração desses modelos em contextos clínicos.

6.1 Trabalhos Futuros

A inteligência artificial como ferramenta para auxílio ao diagnóstico veio para ficar. Os órgãos reguladores e conselhos responsáveis estão a cada dia adaptando mais suas resoluções para permitir a adição de tecnologia no dia a dia das partes interessadas, isso vem acontecendo não só na radiologia odontológica mas na medicina como um todo. Sendo assim as novas técnicas de visão computacional que surgem com propósitos mais genéricos tem nesse setor uma aplicação que deve ser explorada.

Neste estudo, verificou-se que o modelo R-CNN exibiu uma precisão superior em comparação com outros modelos avaliados. Tal superioridade pode ser atribuída às estratégias adotadas e à integração de múltiplas redes neurais para otimizar o desempenho. Diante disso, a investigação de abordagens análogas utilizando modelos de

Vision Transformer é indicada, especialmente considerando que esses modelos demonstraram desempenho superior à CNN quando avaliados de forma isolada.

Pesquisas recentes têm adotado abordagens mais elaboradas na aplicação de transformers. Por exemplo, o trabalho "*End-to-End Object Detection with Transformers*" (CARION et al., 2022) empregou uma rede neural convolucional (CNN) para extração de características da imagem, utilizando posteriormente esse mapa de características como insumo para o modelo transformer, alcançando resultados expressivos na identificação de objetos. Da mesma forma, "ViTPose: *Simple Vision Transformer Baselines for Human Pose Estimation*" (XU et al., 2022) fez uso de redes neurais convolucionais para identificação de objetos, aplicando subsequentemente técnicas de segmentação via transformers para a detecção de pontos de pose em figuras humanas. Os desfechos desses estudos sugerem que a adoção de técnicas semelhantes para a detecção de pontos anatômicos em telerradiografias pode ser vantajosa, sobretudo à luz dos resultados desta pesquisa que destacaram a primazia do modelo R-CNN.

Além disso, pesquisas emergentes buscam redefinir as redes neurais convolucionais para aproximá-las dos transformers. No estudo "*A ConvNet for the 2020s*" (LIU et al., 2022), há uma tentativa de remodelar a estrutura da ResNet, incorporando características análogas às dos transformers. O resultado desse refinamento foi denominado ConvNeXt, e sua eficácia, ao compará-la com transformers, foi evidenciada. Portanto, a aplicação destas ConvNeXt para detecção de pontos anatômicos em telerradiografias também se apresenta como uma linha promissora para investigações futuras.

Outro ponto a ser observado são os avanços relacionados à aquisição de imagem. Nos últimos anos vem se popularizando na odontologia a aquisição de tomografias da região dentomaxilofacial. Essa popularização se deve ao desenvolvimento de uma nova tecnologia de aquisição de imagens chamada cone beam, essa tecnologia permite que a tomografia seja obtida por raios-x disparados de um feixe cônico. Essa nova técnica permite a construção de tomógrafos menores e de menor custo, especialmente indicados para aquisição de imagens da região dentomaxilofacial. O tomógrafo cone beam se assemelha a um aparelho de radiografia panorâmica e opera com o paciente na posição sentada. A aquisição de imagem é mais rápida do que em um aparelho de tomografia convencional e o software necessário para construção da imagem pode ser instalado em

computadores convencionais, dispensando a necessidade de uma workstation como nos aparelhos de tomografia convencional. Outra vantagem do tomógrafo cone beam é que a dose de radiação necessária também é muito menor do que nos aparelhos de tomografias convencionais (GARIB, 2007).

Os primeiros estudos sobre tomografia cone beam foram realizados recentemente na década de 90 e o primeiro tomógrafo cone beam, chamado de Newtom-9000, foi criado em 1998 (GARIB, 2007). Desde então essa tecnologia já se popularizou muito sendo o tomógrafo cone beam um dos aparelhos mais vendidos nos congressos de radiologia odontológica.

Tendo em vista a popularização dos aparelhos de tomografia na odontologia e número crescente de exames realizados sobre os volumes 3D produzidos acredita-se que pesquisas relacionadas à encontrar pontos anatômicos em tomografias também tem relevância para a radiologia odontológica. Além disso, já existem trabalhos acadêmicos aplicando as técnicas de Redes Neurais Convolucionais em volumes com objetivo de reconhecer objetos de maneira similar ao reconhecimento de objetos em imagens. Sendo assim o objetivo de encontrar pontos anatômicos em tomografias odontológicas parece factível.

Ao longo desta pesquisa, identificou-se que um dos principais diferenciais deste trabalho em comparação com outros relacionados reside na ampla base de dados de telerradiografias utilizada. Enquanto os conjuntos de dados apresentados na literatura relacionada não ultrapassaram a marca de 1000 radiografias, esta pesquisa trabalhou com um conjunto de 30.000 imagens, proporcionando uma robustez sem precedentes na área. Dada essa vastidão de dados, um direcionamento promissor para trabalhos futuros seria a estruturação desse conjunto de dados conforme os padrões adotados pelo dataset COCO, especialmente no que tange à anotação de pontos de poses humanas. Adotar tal padrão facilitaria a integração e aplicação de diversas tecnologias emergentes de detecção de pontos, visto que o COCO é uma referência amplamente reconhecida e utilizada para a validação de modelos em estudos científicos. A criação de um banco de dados de imagens de telerradiografias anotadas seguindo o padrão do COCO representaria, indubitavelmente, uma valiosa contribuição para a comunidade científica, potencializando avanços no cruzamento entre a odontologia e a visão computacional.

6.2 Contribuições

Este trabalho de mestrado, ao buscar pontes entre a computação e a odontologia, ilumina um campo de estudo interdisciplinar, promovendo contribuições significativas em várias dimensões. Aqui, delineamos estas contribuições de forma sistemática e mostramos que o escopo desta pesquisa transcende o mero avanço acadêmico. Seus tentáculos se estendem, impactando positivamente o mercado profissional, elevando o padrão de práticas de odontologia e, em última análise, contribuindo para a saúde e bem-estar da população em geral. Ao alinhar rigor científico com aplicabilidade prática, este trabalho de mestrado destaca-se como um marco na confluência da odontologia e da computação.

6.2.1 Contribuições Acadêmicas

Para a Comunidade Científica a presente pesquisa preenche uma lacuna na literatura ao fornecer um comparativo detalhado de tecnologias aplicadas à detecção de pontos em telerradiografias. Isso amplia o corpo de conhecimento e oferece diretrizes para futuros estudos na intersecção entre a odontologia e a computação.

Ao explorar e avaliar técnicas de visão computacional na detecção automática dos pontos anatômicos para exames de traçado cefalométrico, esta pesquisa também contribui para a visão computacional e campos relacionados. Os *insights* derivados podem orientar o refinamento de algoritmos e abordagens, além de evidenciar novos horizontes de aplicação dessas técnicas.

Além disso, no presente estudo empregou-se uma base de dados de telerradiografias com uma quantidade de exemplos superior àquelas reportadas em literaturas correlatas. Tal abordagem evidencia a lacuna existente em relação à disponibilidade de bases extensas para identificação de pontos anatômicos. Em particular, o treinamento realizada durante os experimentos, que evidenciou que a precisão de alguns modelos no conjunto de validação, somente estabiliza após o processamento de mais de 5 mil imagens, reforça a noção de que bases de dados, como a 'Airiz' (KHALID et al., 2023), contendo apenas 1000 imagens, não oferecem um volume adequado para um treinamento eficaz de modelos de detecção de pontos em telerradiografias. Portanto, esta ampliação na quantidade de exemplos configura-se como uma significativa contribuição acadêmica desta pesquisa.

6.2.2 Contribuições Profissionais e Pessoais

A realização deste trabalho proporcionou um profundo enriquecimento acadêmico e técnico ao pesquisador que realizou o trabalho. A natureza interdisciplinar do estudo não apenas consolidou a capacidade de integrar diferentes campos de conhecimento, mas também refletiu na formação profissional, contribuindo para uma visão holística dos desafios encontrados no mercado.

A inserção do pesquisador no mercado amplifica o alcance desta pesquisa. Ao transferir conhecimentos acadêmicos para práticas profissionais, há uma reverberação direta nos processos, na precisão e na eficiência dos exames de traçado cefalométrico, beneficiando, assim, o setor da radiologia odontológica.

6.2.3 Contribuições Sociais

A saúde bucal, sendo crucial para o bem-estar geral da população, é diretamente influenciada por avanços diagnósticos. Ao propor um modelo otimizado para o exame de traçado cefalométrico, esta pesquisa contribui para diagnósticos mais rápidos e precisos, o que leva consequentemente a tratamentos mais eficazes. O reflexo disso é uma sociedade mais saudável e um padrão de saúde bucal elevado.

É notório que muitos profissionais de odontologia têm limitado acesso ou familiaridade com ferramentas computacionais avançadas. Este trabalho, portanto, atende a essa demanda, tornando o acesso e a aplicação destas ferramentas mais tangíveis para os odontologistas, promovendo uma prática clínica mais informada e tecnologicamente avançada.

O escopo desta pesquisa transcende o mero avanço acadêmico. Seus tentáculos se estendem, impactando positivamente o mercado profissional, elevando o padrão de práticas de odontologia e, em última análise, contribuindo para a saúde e bem-estar da população em geral. Ao alinhar rigor científico com aplicabilidade prática, este trabalho de mestrado destaca-se como um marco na confluência da odontologia e da computação.

Referências Bibliográfias

ARMATO III, SAMUEL G., MARYELLEN L. Giger, HEBER MacMahon. 2001. Automated detection of lung nodules in CT scans: Preliminary results. **Radiation imaging physics**, [S.l], v. 28, n. 8, p. 1552-1561, 9 ago. 2001. Disponível em: https://aapm.onlinelibrary.wiley.com/doi/abs/10.1118/1.1387272. Acesso em: 12.11. 2022.

BRASIL. **Lei nº 13.709, de 14 de agosto de 2018**. Lei Geral de Proteção de Dados Pessoais (LGPD). Brasília, DF: Governo Federal, 2018. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm. Acesso em: 13.11. 2022.

BRASIL. **Lei nº 13.787, de 27 de dezembro de 2018**. Dispõe sobre a digitalização e a utilização de sistemas informatizados para a guarda, o armazenamento e o manuseio de prontuário de paciente. Brasília, DF: Governo Federal, 2018. Disponível em: https://legislacao.presidencia.gov.br/atos/?tipo=LEI&numero=13787&ano=2018&ato=5b4ITWE5UeZpWT0d4. Acesso em: 13.11. 2022.

BRASIL. Ministério da Saúde. **PNSB - Política Nacional de Saúde Bucal**. Brasília: Secretaria de Atenção Primária à Saúde, 2003. Disponível em: https://aps.saude.gov.br/politicas/pnsb. Acesso em: 11.11. 2022.

BRASIL. **Resolução CFM nº 1.821, de 11 de julho de 2007**. Aprova as normas técnicas concernentes à digitalização e uso dos sistemas informatizados para a guarda e manuseio dos documentos dos prontuários dos pacientes, autorizando a eliminação do papel e a troca de informação identificada em saúde. Brasília, DF: Governo Federal, 2007. Disponível em: https://www.gov.br/conarq/pt-br/legislacao-arquivistica/resolucoes/resolucao-cfm-no-1-821-de-11-de-julho-de-2007. Acesso em: 15.11. 2022.

BROWN, CHRISTOPHER M., DANA H. Ballard. **Computer vision**. [S.l]: Prentice-Hall, 1982. 539 p.

CALAS, GREGORIO Maria Julia, GUTFILEN Bianca, PEREIRA, Wagner C. de A. CAD e mamografia: por que usar esta ferramenta? **Radiologia Brasileira**, [S. l], v. 45, n. 1, fev. 2012.

CARION, Nicolas, MASSA, Francisco, SYNNAEVE, Gabriel, USUNIER, Nicolas, KIRILLOV, Alexander, ZAGORUYKO, Sergey. **End-to-End Object Detection with Transformers**. [S. l]: arXiv. 2022. Disponível em: https://arxiv.org/abs/2005.12872. Acesso em: 15.09. 2023.

CFAZ.NET. **Software de Cefalometria Computadorizada**. 2020.

CFO. **Quantidade Geral de Cirurgiões-Dentistas Especialistas**. CFO, 2022. Disponível em: https://website.cfo.org.br/estatisticas/quantidade-geral-de-cirurgioes-dentistas-especialistas/. Acesso em: 15.11. 2022.

CHEN, Michael Y. M., POPE, Thomas L., OTT, David J. **Radiologia Básica**. Espanha: AMGH, 2014. 428 p.

CWUR. Center for World University Rankings. **Rankings by Subject - 2017**. 2017. Disponível em: https://cwur.org/2017/subjects.php#Dentistry,%20Oral%20Surgery%20&%20Medicine. Acesso em: 13.11. 2022.

DAIMLER, AG. **Semantic Understanding of Urban Street Scenes**. Cityscapes Dataset, 2016. Disponível em: https://www.cityscapes-dataset.com/. Acesso em: 16.11. 2022.

DALAL, N., TRIGGS. **Histograms of oriented gradients for human detection.** IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), p. 886-893, 2005.

DOI, Kunio. Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. **Computerized Medical Imaging and Graphics**, v. 31, n. 4-5, p. 198-211, jul. 2007.

DOSOVITSKIY, Alexey. **An Image is Worth 16x16 Words**: Transformers for Image Recognition at Scale. [S. l]: arXiv. 2020. Disponível em: https://doi.org/10.48550/arXiv.2010.11929. Acesso em: 14.11. 2022.

FELZENSZWALB, P. F., GIRSHICK, R. B., MCALLESTER, D. **Cascade object detection with deformable part models**. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), p. 2241-2248, 2010.

FREITAS, Aguinaldo D. **Radiologia odontológica**. [S. l]: Artes Médicas, Divisão Odontológica, 2004. 833 p.

FRISBY, J. P., STONE James. V. **The computational approach to biological vision**. Seeing. Londres: MIT Press, 2 ed., 2010, 30 p. Disponível em: http://mitp-content-server.mit.edu:18180/books/content/sectbyfn?collid=books_pres_0&id=8574&fn=9780262514279_sch_0001.pdf. Acesso em: 15.11. 2022.

GARIB, Daniela G. Tomografia computadorizada de feixe cônico (Cone beam): entendendo este novo método de diagnóstico por imagem com promissora aplicabilidade

na Ortodontia. **Rev. Dent. Press Ortodon**. Ortop. Facial, v. 12, n. 2, abr. 2007. Disponível em: https://doi.org/10.1590/S1415-54192007000200018. Acesso em: 23.11. 2022.

GOODFELLOW, Ian J. **Generative Adversarial Networks.** [S. l]: arXiv. 2014. Disponível em: https://arxiv.org/abs/1406.2661. Acesso em: 13.11. 2022.

HE, Kaiming, GKIOXARI Georgia, DOLLÁR Piotr, GIRSHICK Ross. **Mask R-CNN**. [S. l]: arXiv. 2017. Disponível em: https://arxiv.org/abs/1703.06870. Acesso em: 14.11. 2022.

HUBEL, David H., WIESEL, Torsten N. Effects of monocular deprivation in kittens. **Naunyn Schmiedebergs Arch Exp Pathol Pharmakol**, v. 248, p. 492-497, 1964.

IDC. **Worldwide IT Industry 2021 Predictions and Latin America Implications**: Building Resiliency to Thrive in the Next Normal. IDC, 2020. Disponível em: http://www.idclatin.com/2020/webinars/9_nov_fe/FutureScape2021_LatAm.pdf. Acesso em: 12.11. 2022.

INTERLANDI, S. **O cefalograma padrão do curso de pós-graduação de Ortodontia da Faculdade de Ortodontia da Faculdade de Odontologia da USP.** Rev. Fac. Odontol. S. Paulo, v. 6, n.1, p. 63-74 jan/mar, 1968.

KANG, Dae-Young, DUONG, Pham, PARK, Jung-Chul. Application of Deep Learning in Dentistry and Implantology. **The Korean Academy of Oral and Maxillofacial Implantology**. v. 24, p.148-181, 2020.

KHALID, Muhammad Anwaar, ZULFIGAR, Kanwal, BASHIR, Ulfat, SHAHEEN, Areeba, IQBAL, Rida, RIZWAN, Zarnab, RIZWAN, Ghina, FRAZ, Muhammad Moazam. **'Aariz: A Benchmark Dataset for Automatic Cephalometric Landmark Detection and CVM Stage Classification.** [S. l]: arXiv. 2023. Disponível em: https://arxiv.org/abs/2302.07797. Acesso em: 15.10. 2023.

LANGLAND, Olaf E., LANGLAIS, Robert P. **Principles of Dental Imaging**. [S.l.]: Lippincott Williams & Wilkins.1997.

LETO, Daniel N. **Métodos de processamento de imagens para Visão Computacional no Cubo de Rubric**. 2015. Trabalho de Conclusão de Curso (Bacharelado em Engenharia Elétrica) – Curso de Engenharia Elétrica, CEFET-MG, Belo Horizonte, 2015. Disponível em: https://www2.dee.cefetmg.br/wp-content/uploads/sites/18/2017/12/TCC_2015_2_DLNCosta.pdf. Acesso em: 23.11. 2022.

LI, Stan Z., JAIN, Anil K. **Handbook of Face Recognition**. [S.l]: Springer London, 2011.

Lin, T. **Common Objects in Context (COCO).** 2015. Disponível em: https://cocodataset.org/. Acesso em: 24.11. 2022.

LIU, Zhuang, MAO, Hanzi, WU, Chao-Yuan, FEICHTENHOFER, Christoph, DARRELL, Trevor, XIE, Saining. **A ConvNet for the 2020s**. [S. l]: arXiv. 2022. Disponível em: https://arxiv.org/abs/2201.03545. Acesso em: 15.09. 2023.

LIVINGSTONE, Margaret. **Vision and Art**: The Biology of Seeing. [S.l]: Harry N. Abrams, 2008. 208 p.

LOWE, David G. Three-Dimensional Object Recognition from Single Two-Dimensional Images. **Artificial Intelligence**, p. 355–395, mar. 1987. Disponível em: https://www.cs.ubc.ca/~lowe/papers/aij87.pdf. Acesso em: 13.11. 2022.

MARR, David. **Vision**: A Computational Investigation Into the Human Representation and Processing of Visual Information. [S.l]: W.H. Freeman,1982.

MARTINS, Wilson D. **Wilhelm Conrad Rötgen e a descoberta dos raios X**. 2005. Disponível em: http://www.imaginologia.com.br/. Acesso em: 24.11. 2022.

MINS. **Stanford**, 2020. Disponível em: https://m.blog.naver.com/khm159/221832329861. Acesso em: 23.11. 2022.

MUSEUM DEUTSCHES RONTGEN. **Development of X-rays.** [2022]

NAIMISH, Agarwal. 2017. **Facial Key Points Detection using Deep Convolutional Neural Network - NaimishNet**. [S. l]: arXiv. 2017. Disponível em: https://arxiv.org/abs/1710.00977. Acesso em: 13.11. 2022.

ONU. **Declaração Universal dos Direitos Humanos**. UNICEF, 1948. Disponível em: https://www.unicef.org/brazil/declaracao-universal-dos-direitos-humanos. Acesso em: 21.11. 2022.

OPENAI. **Dall.2**. [200?]. Disponível em: https://openai.com/dall-e-2/. Acesso em: 21.11. 2022.

PAIVA, Helson J. **Noções e Conceitos Básicos em Oclusão, Disfunção Temporomandibular e Dor Orofacial**. 1 ed. Curitiba: Santos, 2008.

PALMER, Stephen E. **Vision Science**: From Photons to Phenomenology. Londres: MIT Press, 2010.

PINKER, Steven. **Visual Cognition (MIT Press)**. [S. l.]: Bradford Books, 1986.

QUEK, ST., THNG, C. H., KHOO, J. B. K., KOH, W. L. Radiologists detection of mammographic abnormalities with and without a computer-aided detection system. **Australasian Radiology**, v. 47, p. 257-260, ago. 2003. Disponível em:

https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1440-1673.2003.01173.x. Acesso em: 25.11.. 2022.

ROBERTS, Lawrence G. **Machine Perception of Three-dimensional Solids**. [S.l]: Garland Pub, 1980.

SALU, Enio J. **Geografia Econômica da Saúde no Brasil**. 2020. Disponível em: http://www.gesb.com.br/index_arquivos/Page316.htm. Acesso em: 25.11. 2022.

SHI, Jianbo, MALIK, Jitendra. **Normalized Cuts and Image Segmentation**. IEEE Transactions on Pattern Analysis and Machine Intelligence, v. 22, n. 8, ago. 2000. Disponível em: https://people.eecs.berkeley.edu/~malik/papers/SM-ncut.pdf. Acesso em: 11.11. 2022.

SHU, Michelle. **What machine learning borrowed from vision Science**. Médium, ago. 2013. Disponível em: https://medium.com/@michellewshu. Acesso em: 13.11. 2022.

SONG, Yu, QIAO, Xu, IWAMOTO, Yutaro. **Automatic Cephalometric Landmark Detection on X-ray Images Using a Deep-Learning Method. Applied Sciences**. v. 10, abr. 2020. Disponível em: 10.3390/app10072547. Acesso em: 14.11. 2022.

SZELISKI, Richard. **Computer Vision**: Algorithms and Applications. [S. l]: Springer International Publishing, 2022.

VASWANI, Ashish. **Attention Is All You Need**. [S. l]: arXiv. 2017. Disponível em: https://arxiv.org/abs/1706.03762. Acesso em: 12.11. 2022.

VILELLA, Oswaldo D. **Manual de cefalometria**. [S. l]: Revinter, 2009.

VIOLA, P., JONES, M. **Rapid object detection using a boosted cascade of simple features**. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), p.1, dez. 2001. Disponível em: 10.1109/CVPR.2001.990517. Acesso em: 13.11. 2022.

XU, Yufei, ZHANG, Jing, ZHANG Qiming, TAO, Dacheng. **ViTPose: Simple Vision Transformer Baselines for Human Pose Estimation**. [S. l]: arXiv. 2022. Disponível em: https://arxiv.org/abs/2204.12484. Acesso em: 15.09. 2023.

WANDELL, B. A. **Foundations of Vision**.1995. Disponível em: https://foundationsofvision.stanford.edu/. Acesso em: 14.11. 2022.

WIKIMEDIA COMMONS... Disponível em:https://commons.wikimedia.org. Acesso em: 11.11. 2022.