UNIVERSIDADE DO SUL E SUDESTE DO PARÁ

INSTITUTO DE GEOCIENCIA E ENGENHARIAS

Faculdade de Engenharia da Computação

ANÁLISE DO VIÉS EQUATORIAL EM VÍDEOS 360°: IMPACTO DO ÁUDIO NA DISTRIBUIÇÃO DAS FIXAÇÕES OCULARES.

CARLENO CÁSSIO SANTOS ASSUNÇÃO

Marabá - PA

2025

CARLENO CÁSSIO SANTOS ASSUNÇÃO

ANÁLISE DO VIÉS EQUATORIAL EM VÍDEOS 360°: IMPACTO DO ÁUDIO NA DISTRIBUIÇÃO DAS FIXAÇÕES OCULARES.

Projeto Final de Curso II apresentado à Universidade Federal do Sul e Sudeste do Pará, como parte dos requisitos necessários para obtenção do Título de Bacharel em Engenharia da Computação.

Orientadora: Prof.ª Aline Farias Gomes de Sousa

Marabá - PA

2025

CARLENO CÁSSIO SANTOS ASSUNÇÃO

"ANÁLISE DO VIÉS EQUATORIAL EM VÍDEOS 360°: IMPACTO DO ÁUDIO NA DISTRIBUIÇÃO DAS FIXAÇÕES OCULARES"

Projeto Final de Curso II apresentado à Universidade Federal do Sul e Sudeste do Pará, como parte dos requisitos necessários para obtenção do Título de Bacharel em Engenharia da Computação.

Marabá: 07 de março de 2025.

Banca:

Aline Farias Gomes de Sousa

Profª. Me. Aline Farias Gomes de Sousa - Orientadora

Jeânderson de Melo Dantas

Profº. Drº. Jeanderson de Melo Dantas

Erberson Rodrigues Pinheiro dos Santos

Profº. Me. Erberson Rodrigues Pinheiro dos Santos

Marabá - PA.

2025

Este trabalho é todo dedicado à minha esposa Viviane Holanda, meus filhos, à minha mãe Conceição, à minha irmã Renata e ao meu avô Nazaré Bispo, pois é graças a eles que hoje posso concluir o meu curso.

AGRADECIMENTOS

Quero agradecer, em primeiro lugar, a DEUS, por ter me dado forças e coragem para enfrentar esse grande desafio que foi conquistar minha primeira graduação, de várias que ainda virão. Quero agradecer também a toda a minha família, pelo apoio incondicional, pois foram eles que sempre estiveram ao meu lado.

À minha mãe, Maria da Conceição Bispo dos Santos, que sempre me incentivou. À minha esposa, Viviane da Silva Holanda, que esteve ao meu lado em todos os momentos e nunca me deixou desanimar. Mesmo diante das dificuldades, sempre me apoiou.

Aos meus filhos, Fernando Garcia Ferreira Assunção, Calyeh Victor Holanda Assunção e Caylah Maria Holanda Assunção, e também à nossa pet, Lollah Maria Holanda Assunção, que faz parte da família, pois sem eles nada disso estaria acontecendo.

Agradeço também a todos os professores da Faculdade de Engenharia da Computação – FEC, que, com muita sabedoria, souberam ensinar não só a mim, mas a todos os alunos que já se formaram e que ainda estão se formando nesta instituição. Durante esses anos, aprendi muito e conquistei amizades maravilhosas.

Por isso, deixo aqui meu mais sincero agradecimento. Obrigado!

RESUMO

Vídeos 360° proporcionam experiências imersivas ao permitir que o espectador explore livremente o ambiente ao seu redor, utilizando movimentos da cabeça e dos olhos. No entanto, estudos indicam a existência de um viés equatorial, no qual as fixações oculares tendem a se concentrar na região central da esfera visual. A integração de estímulos auditivos pode influenciar significativamente a distribuição das fixações, potencialmente reduzindo esse viés ou reforçando padrões de atenção específicos. Este trabalho investiga como a presença ou ausência de áudio afeta a distribuição das fixações em vídeos 360°, utilizando uma base de dados de rastreamento ocular. Para comparar as distribuições das fixações, foram aplicados testes estatísticos t (para médias) e F (para variâncias), a fim de avaliar diferenças significativas entre as condições com e sem áudio. Os resultados contribuem para uma melhor compreensão do impacto multimodal na atenção visual, com aplicações em realidade virtual, entretenimento e educação, além de oferecer subsídios para o desenvolvimento de modelos computacionais de saliência mais precisos.

Palavras-chave: Realidade Virtual (RV), Vídeos 360°, Atenção Visual, Viés Equatorial, Mapas de Saliência, Áudio, Fixações Oculares.

ABSTRACT

360° videos provide immersive experiences by allowing the viewer to freely explore the environment around them using head and eye movements. However, studies indicate the existence of an equatorial bias, in which eye fixations tend to concentrate in the central region of the visual sphere. The integration of auditory stimuli can significantly influence the distribution of fixations, potentially reducing this bias or reinforcing specific attentional patterns. This work investigates how the presence or absence of audio affects the distribution of fixations in 360° videos, using an eye tracking database. To compare the distributions of fixations, t (for means) and F (for variances) statistical tests were applied in order to assess significant differences between the conditions with and without audio The results contribute to a better understanding of the multimodal impact on visual attention, with applications in virtual reality, entertainment and education, in addition to offering support for the development of more accurate computational saliency models.

Keywords: *Virtual Reality (VR), 360° Videos, Visual Attention, Equatorial Bias, Saliency Maps, Audio, Eye Fixations.*

LISTA DE ABREVIATURAS E SIGLAS

2D - Bidimensional (*Two-Dimensional*)

EM - Movimento dos Olhos (*Eye Movement*)

CNN - Redes Convolucionais (Convolutional Neural Networks)

HMDs - Visores montados na cabeça (Head-Mounted Displays)

HM - Movimento da Cabeça (*Head Movement*)

ITU - União Internacional de Telecomunicações (*International Telecommunications Union*)

RoI - Região de Interesse (*Region of Interest*)

RV - Realidade Virtual (*Virtual Reality*)

VQA - Avaliação de Qualidade Visual (*Visual Quality Assessment*)

1. INTRODUÇÃO

Atualmente, observa-se um rápido desenvolvimento da tecnologia de realidade virtual (RV) (LI et al., 2019), com os vídeos 360° emergindo como um tipo essencial de conteúdo imersivo. Diferente dos vídeos bidimensionais (2D), os vídeos 360° permitem ao espectador explorar livremente uma cena em um campo de 360 graus, proporcionando uma experiência altamente envolvente. Essa liberdade de exploração visual é acompanhada de desafios técnicos, como a alta resolução e a necessidade de compressão eficiente para transmissão e armazenamento, o que impulsiona pesquisas voltadas para a modelagem da atenção visual.

Para cobrir toda essa faixa de visualização com alta fidelidade, a resolução de vídeo 360° deve ser extraordinariamente alta, impondo pesadas cargas aos sistemas de armazenamento e transmissão desses vídeos.

Nas últimas décadas, muitos padrões de compressão de vídeo/imagem foram desenvolvidos para formatos 2D tradicionais pela União Internacional de Telecomunicações (ITU - *International Telecommunications Union*). Contudo, devido às características esféricas dos vídeos imersivos, as abordagens 2D para armazenar, processar e transmitir vídeos (CHEN; LI; ZHANG, 2018) mostraram-se sub-ótimas para vídeo 360°. A fim de melhorar a eficiência na representação e exibição da informação, a comunidade científica tem investido em novas estratégias de codificação específicas para esse tipo de vídeo (XU et al., 2018; NGUYEN; YAN; NAHRSTEDT, 2018; CHENG et al., 2018) e assim aliviar as cargas de armazenamento e transmissão, o que exige um processamento dedicado e mais eficiente para vídeo imersivos.

Para medir o desempenho de um esquema de codificação é necessário avaliar a

VQA (*visual quality assessment*), ou seja, a qualidade visual da imagem decodificada e a taxa de compressão associada (LI et al., 2019). Naturalmente, surgiram vários trabalhos sobre a VQA em vídeo 360°, tanto a partir da coleta efetiva de dados subjetivos de qualidade (Upenik; Řeřábek; Ebrahimi, 2016; Xu et al., 2017), quanto mediante modelagem da qualidade visual (CHEN; LI; ZHANG, 2018).

Segundo Li et. al. (2019), a visualização de vídeo 360° através da *viewport*, que corresponde a porção de área visível, explica dois fatos:

1. A degradação da qualidade na janela de visualização é mais visível, já que o espectador se concentra apenas na *viewport*.
2. Há uma redundância em massa dos bits codificados de vídeo de 360°, já que a grande região fora da *viewport* é invisível para o espectador.

Inspirada nesses fatos, considerar elementos da percepção visual humana pode beneficiar o compromisso entre qualidade, complexidade computacional e taxa de compressão em esquemas de codificação de vídeo 360°. Há também trabalhos concentrando-se na modelagem de atenção visual para esse tipo de mídia para predizer as regiões com maior percentual de atrair o olhar de um observador (GUTIéRREZ et al., 2018).

Muitos trabalhos relacionados ao VQA de vídeos 360° continuam a ser desenvolvidos e isso tem permitido muitos avanços em Modelagem de Atenção Visual. Como a maioria dos trabalhos existentes se concentra na extração de características espaciais para o VQA em vídeos 360°, a extração de características espaço-temporais para VQA em vídeo 360º se mostra como uma tendência promissora (LI et al., 2019).

O movimento combinado de cabeça (HM, *head movement*) e olhos (EM, *eye*

movement) dos usuários definem tanto o *viewport* como a região dentro do *viewport* a ser analisada com maiores detalhes (ou seja, porção da imagem visualizada que incide na fóvea[1]). Em particular, o HM refere-se às localizações dos *viewports* dos sujeitos, enquanto o EM reflete a região de interesse (RoI, *region of interest*) dentro dos *viewports*.

Para realizar a previsão de atenção visual em vídeos esféricos é necessário prever o HM e o EM. Esse mecanismo de previsão de atenção visual em vídeos 2D não é necessário, importando apenas o EM. Pode-se afirmar então que a necessidade de estimar o HM é uma das principais diferenças entre os modelos de atenção visual de vídeos 360° e vídeos 2D (LI et al., 2019).

Os dados obtidos a partir do EM são representados através de movimentos oculares, como fixação e sacadas. O movimento ocular fornece uma imagem clara do comportamento de busca e foco de um participante. Para a aquisição de dados do movimento ocular, são utilizados equipamentos de rastreamento ocular, ou seja, *eye/gaze-trackers*.

Com base nos dados resultantes de rastreamento ocular e do movimento da cabeça, torna-se possível prever a atenção visual humana por meio de modelos computacionais.

Estes modelos computacionais podem gerar mapas de saliência, que é um mapeamento que atribui uma importância a dado objeto ou região da imagem/vídeo. É importante salientar que nesta pesquisa mapas saliência serão conceituados como um mapeamento gerado por modelos computacionais, enquanto que mapas de fixação serão conceituados como um mapeamento proveniente das fixações observadas registradas por um *eye/gaze-trackers*.

---

1. Fóvea: Região central da retina humana responsável pela maior acuidade visual, devido à alta densidade de células fotorreceptoras do tipo cone. Em modelos de atenção visual, a fóvea é frequentemente considerada ao simular o comportamento do olhar humano, influenciando a alocação de atenção e a percepção de detalhes em imagens e vídeos.

Os mapas de saliência possuem inspiração biológica, ou seja, são inspirados em resultados de experimentos sobre a percepção visual humana, percepção essa que seleciona as áreas mais importantes de uma imagem/vídeo para fixação do olhar. Como afirma Driver, J. (2001), estudos de atenção visual humana comprovam que o cérebro humano seleciona regiões de uma imagem/vídeo para fixar o olhar.

Esta escolha humana tem dois modos de funcionamento, o modo *top-down* e o modo *bottom-up*. No modo *top-down*, a atenção é guiada por fatores internos, como conhecimento prévio, expectativas e objetivos do observador. Esse processo é deliberado e influenciado por experiências anteriores. Já no modo *bottom-up*, a atenção é capturada por estímulos externos, como contrastes visuais, cores vibrantes ou movimentos inesperados. Esse processo é automático e impulsionado pelas características da cena.

O que acontece é que os mapas de saliência gerados por computador ainda não conseguem prever suficientemente o comportamento do olhar humano e, portanto, são necessárias melhorias na geração de mapas de saliência. Com o uso de *Deep Learning* é possível que melhorias sejam feitas, já que a construção de grandes bancos de dados (*Big Data*) de mapas de saliência é algo inerente ao processo de previsão do comportamento do olhar humano.

Deep learning é um tipo de *machine learning* que tenta imitar a rede neural do cérebro humano, capaz de treinar a máquina para realizar tarefas de maneira mais natural. Esse treinamento é feito de forma que a máquina possa aprender sozinha através do reconhecimento de padrões em várias camadas de processamento.

Um exemplo do uso de *Deep Learning* são as Redes Neurais Convolucionais (CNN - do inglês *Convolutional Networks Neurals*). Uma CNN tende a demandar um nível

mínimo de pré-processamento quando comparada a outros algoritmos de classificação de imagens (Lecun; Bottou; Bengio & Haffner, 1998).

A modelagem da atenção visual é essencial para identificar as regiões de maior interesse em uma cena, permitindo otimizar recursos computacionais e melhorar a experiência do usuário. Além disso, a integração de estímulos multimodais, como o áudio, pode influenciar significativamente os padrões de atenção. Estudos indicam que o áudio não apenas complementa a experiência visual, mas também guia o olhar para regiões específicas, reduzindo a dispersão das fixações. Entretanto, a influência do áudio sobre o viés equatorial em vídeos 360° ainda é uma questão em aberto.

Neste trabalho, busca-se investigar como a presença ou ausência de áudio afeta a distribuição das fixações visuais de observadores ao assistir vídeos 360°, utilizando dados de rastreamento ocular. A compreensão dessa dinâmica pode contribuir tanto para o desenvolvimento de modelos computacionais mais precisos quanto para aplicações práticas, como realidade virtual e sistemas de entretenimento.

Este trabalho está estruturado nos seguintes capítulos:

O capítulo 1 aborda uma introdução sobre o presente trabalho, demonstrando os seus conceitos, objetivos e motivações do mesmo. No capítulo 2 é apresentada a revisão de literatura, empregando os principais conceitos da temática de atenção visual em vídeos imersivos, os modelos computacionais de atenção visual, assim como o impacto do áudio na atenção visual. No capítulo 3 são apresentados os materiais e métodos empregados no decorrer deste trabalho, assim como o planejamento do mesmo. Já no capítulo 4, estão os resultados alcançados e as discussões. E por fim, no capítulo 5, apresentamos as considerações finais desde trabalho.

1.1 Justificativa

Os seres humanos são criaturas multissensoriais inteligentes, capazes de detectar e focar um estímulo visual ou de áudio específico em um ambiente desordenado, ou seja, ter comportamento atencional.

No mundo real, as informações recebidas são geralmente simultâneas através de dois ou mais sentidos (audição e visão, por exemplo), o cérebro então faz a junção desses dados para então produzir uma única mensagem coerente (FRATER; ARNOLD; VAHEDIAN, 2001). O áudio é, assim como a imagem, componente inerente aos vídeos. Trabalhos como o de Frater et. al. (2001) demonstrou que a presença de áudio tem um impacto significativo na qualidade subjetiva do vídeo 2D.

No contexto dos vídeos 360°, onde a exploração visual é altamente dinâmica, a presença de áudio pode atuar como um guia atencional, modulando a distribuição das fixações oculares. Estudos demonstram que o áudio influencia significativamente a percepção visual, podendo reforçar padrões de atenção e alterar o viés equatorial.

O viés equatorial refere-se à tendência dos espectadores de fixarem predominantemente suas atenções na região central da esfera visual, limitando a exploração de outras áreas do campo imersivo. Compreender esse viés é essencial para melhorar a apresentação de conteúdos e otimizar a experiência dos usuários em ambientes imersivos.

Trabalhos anteriores mostram que modelos de saliência visual para vídeos 360° frequentemente ignoram o impacto do áudio, comprometendo sua precisão na previsão dos pontos de fixação. Xu et. al. (2018) sugere como temática a ser abordada na área de geração de mapas de saliência em vídeos imersivos, a verificação da influência do áudio.

A integração de estímulos multimodais, como o áudio, tem se mostrado uma

ferramenta poderosa na orientação da atenção visual. Estudos indicam que o áudio pode atuar como um guia, direcionando o olhar para regiões específicas da cena, o que pode resultar na redução do viés equatorial ou no reforço de padrões atencionais. No entanto, a influência do áudio em cenários imersivos ainda é pouco explorada, especialmente em vídeos 360°, onde o som pode ser direcional e contextual.

Melhorias na compreensão do viés equatorial e na geração de mapas de saliência podem impactar diretamente aplicações práticas, como a compressão de vídeos imersivos. Se for possível prever com maior precisão as regiões que atraem fixações, torna-se viável otimizar a qualidade visual dessas áreas e reduzir a resolução em regiões periféricas, aumentando a eficiência da transmissão e armazenamento de vídeos 360°.

Diante do que foi exposto, surge o seguinte problema de pesquisa: Como o áudio afeta a dinâmica da atenção visual e a manifestação do viés equatorial em vídeos 360°?

Além de contribuir para o avanço do conhecimento científico na área, os resultados deste estudo podem ter aplicações práticas em diversas indústrias, como entretenimento, educação e publicidade. Por exemplo, produtores de conteúdo podem usar essas informações para criar experiências mais imersivas e direcionadas, enquanto engenheiros podem desenvolver algoritmos de compressão e transmissão mais eficientes, focados em regiões de maior interesse visual.

Assim, a presente pesquisa busca preencher lacunas existentes na literatura, explorando a interação entre áudio e atenção visual em um contexto imersivo e contribuindo para o desenvolvimento de tecnologias e aplicações inovadoras no campo da realidade virtual.

1.2 Objetivos

O objetivo geral deste trabalho é: Investigar o impacto do áudio na distribuição das fixações oculares e na manifestação do viés equatorial em vídeos.

Os objetivos específicos são:

- Analisar a distribuição das fixações visuais ao longo do eixo equatorial em vídeos 360° com e sem áudio;
- Comparar os mapas de saliência visual gerados nas duas condições;
- Identificar tendências comportamentais na distribuição das fixações em função da presença do áudio.

2. REVISÃO BIBLIOGRÁFICA

2.1 Atenção Visual em Vídeos 360°

O estudo da visão ou atenção visual tem sido feita à décadas, tendo seu princípio nas pesquisas de Yarbus (TATLER, 2010). As pesquisas de Yarbus permitiram gravações estáveis da posição dos olhos durante longos períodos de gravação. O russo desenvolveu dispositivos que permitiam a apresentação de imagens que se moviam com os olhos, de modo que uma imagem retiniana estabilizada pudesse ser apresentada, no entanto como o estudo se limitava apenas a observações oculares e introspecções devido as limitações tecnológicas existente na época, o rastreamento ainda estava limitado a pequenas áreas (TATLER, 2010). Com os avanços tecnológicos, novos testes e observações foram realizados, o que permitiu a ampliação do uso do rastreamento ocular em diversas áreas de pesquisa. Isso tornou o tema interdisciplinar, abrangendo campos como psicofísica, neurociência cognitiva e ciência da computação.

Logo, o rastreamento da trajetória ocular, ou *eye tracking*, é o processo de identificar o olho de uma pessoa e rastrear as movimentações dos olhos. Em geral, a movimentação ocular em primatas pode ser resumida na combinação de quatro movimentos básicos: sacadas, fixações, perseguições suaves e *Nystagmus* (DUCHOWSKI, 2007).

Duchowski (2007) define cada técnica de visualização como:

- Sacadas: Os movimentos sacádicos são voluntários e reflexivos, são considerados desejos voluntários de mudar o foco da atenção.
- Fixações: As fixações são caracterizadas pelos movimentos oculares em

miniatura, como tremor, deriva e microsacadas. São ditas como comportamentos nos quais sozinhos permanecem estacionários em algum aspecto do ambiente.

- Perseguições suaves: As perseguições dependem da amplitude do movimento do alvo e podem ocorrer em qualquer meridiano. São dados como movimentos lentos e contínuos.
- Nystagmus: Atuam como uma combinação de diversos pequenos movimentos. É um mecanismo acoplado dos olhos que tem a função de estabilizar os olhos e garantir uma visão nítida.

A atenção visual em vídeos 360° é um processo complexo influenciado por múltiplos fatores, como a dinâmica da cena, o contexto e os estímulos auditivos. Diferente de vídeos 2D, vídeos imersivos permitem ao espectador explorar livremente o ambiente, tornando fundamental a compreensão dos padrões de fixação ocular e sua relação com o movimento da cabeça (Head Movement - HM).

Segundo Duchowski (2017), a visão pode ser modelada como um processo cíclico composto por três etapas: percepção periférica inicial, realocação da atenção e fixação da fóvea na região de interesse.

2.2 Modelos computacionais de atenção visual em vídeos imersivos

Os modelos de saliência visual são ferramentas computacionais que simulam a percepção humana, prevendo regiões da cena mais prováveis de atrair o olhar. Esses modelos são amplamente utilizados para gerar mapas de saliência, que representam graficamente as áreas de maior interesse visual.

Pesquisas recentes apontam para a relevância da integração de informações

multimodais, como o áudio, para aprimorar a acurácia desses modelos em vídeos 360°.

Vídeos 360° é um tipo de multimídia que fornece ao expectador uma experiência imersiva. Os vídeos "tradicionais" bidimensionais (2D) são bastante diferentes dos vídeos 360°, isto porque os vídeos bidimensionais são limitados a apenas um plano.

Esse tipo de mídia, permite ao expetador selecionar a *viewport* para se concentrar no conteúdo que atrai sua atenção. Para a seleção do *viewport* é utilizando o movimento da cabeça, enquanto que o movimento ocular é quem vai determinar qual a região será capturada em alta resolução (dentro da *viewport*). Uma das principais diferenças entre os modelos de atenção para vídeos 360° e para vídeos 2D é a necessidade de realizar a previsão do movimento da cabeça e do movimento ocular para que seja possível realizar a previsão de atenção visual em vídeos 360° (LI et al., 2019).

Para que seja possível coletar dados do movimento da cabeça e do movimento dos olhos são utilizados dispositivos desenvolvidos para esta finalidade. Com o avanço das tecnologias, é possível verificar o aumento dos tipos de equipamentos disponíveis, como por exemplo, os *Head-Mounted Displays* (HMDs), que utilizam sensores de rastreamento de cabeça para permitir que o espectador ajuste a *viewport* apenas movendo a cabeça, sem a necessidade de controle manual (Figura 1). Esse tipo de mídia é bem realista e permite às pessoas uma sensação muito próxima do que é sentido no mundo real.

Todo esse avanço tecnológico têm salientado ainda mais a necessidade de que se realize um processamento cada vez mais eficaz. O armazenamento e transmissão de vídeos 360° são altos, já que a resolução é extremamente alta. Além disso, é necessário que a taxa de *frames* seja alta, para que se evite enjoo nos expectadores (CHEN; LI; ZHANG, 2018).

Figura 1: Exemplo de *Head-Mounted Displays*

Diante deste contexto, é interessante salientar que existe uma região considerável fora da janela de visualização, região que se torna invisível ao expectador. Isso impacta diretamente na redundância dos bits codificados. Se for possível determinar qual a região que fica invisível ao expectador durante a visualização de conteúdo em vídeos 360°, ou seja, se for possível prever a percepção humana, seria possível realizar a compressão de maneira mais eficiente e beneficiar a VQA.

O Trabalho de Li et. al. (2019) descreve o estado da arte para vídeos/imagem 360º e comenta alguns modelos de atenção visual para esse tipo de mídia. Além disso, cita as bases de dados utilizadas por cada um dos trabalhos elencados. As bases de dados são públicas e estão disponíveis para *download*.

A base de dados Salient360 é bastante utilizada por pesquisadores da área de modelagem de atenção visual. Os conjuntos de dados Salient360 contêm os dados HM e EM para vídeo em 360° (DAVID et al., 2018). Esse repositório possui 19 sequências

de vídeo com resolução de 3840 × 1920 pixels, classificados por 3 grupos. Cada sequência tem duração de 20 segundos e 57 indivíduos visualizaram livremente todas as sequências do experimento. É importante ressaltar que os indivíduos estavam sentados em uma cadeira giratória, quando foram submetidos ao experimento. Para a obtenção dos dados foi utilizado um dispositivo que incorpora não apenas os dados do movimento da cabeça, mas também os dados dos movimentos dos olhos dos indivíduos.

No trabalho de David et. al. (2018) os observadores foram instruídos a explorar livremente vídeos em 360° usando um fone de ouvido. Apesar do uso do fone de ouvido, dos 19 vídeos explorados, apenas 5 foram reproduzidos com áudio. Como para o desenvolvimento do modelo de atenção audiovisual proposto por esta pesquisa envolve características de áudio, se faz necessário uma base de dados que contenha vídeos 360° com som. Outra base de dados que possui estímulos de vídeos 360º está disponível para download em *https://github.com/xuyanyu-shh/VR-EyeTracking* (Xu et al., 2018).

2.3 Impacto do Áudio na Atenção Visual

Estímulos auditivos podem influenciar significativamente a atenção visual, funcionando como guias que direcionam o olhar para regiões específicas da cena. Estudos em vídeos 2D demonstraram que a presença de áudio reduz a dispersão das fixações, aumentando a coerência grupal entre os observadores.

No entanto, a integração do áudio em modelos de atenção visual para vídeos 360° ainda é um campo em desenvolvimento, com desafios como a modelagem da influência espacial do som e sua interação com os elementos visuais da cena.

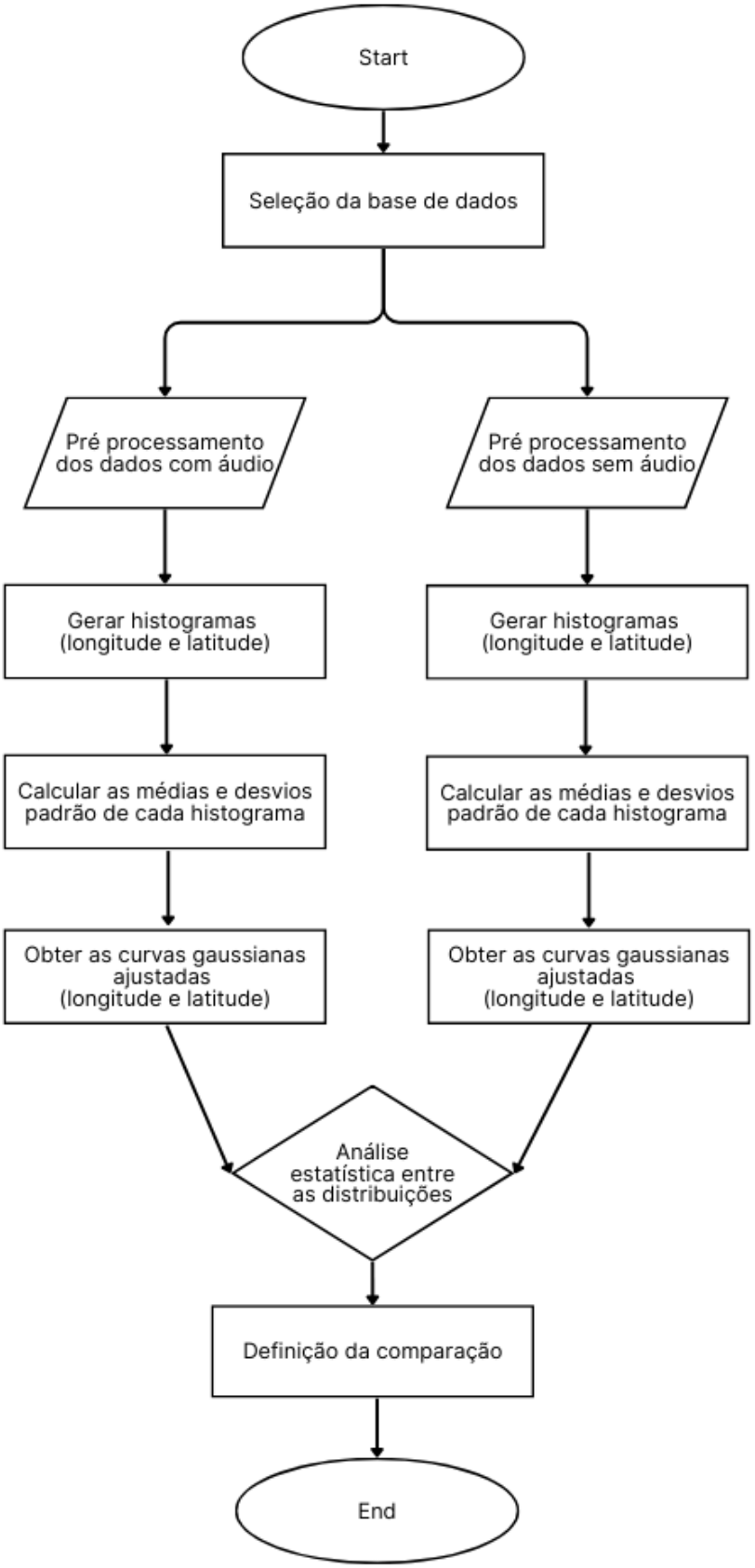
3. MATERIAIS E METODOS

Neste capítulo serãao apresentados os materiais e métodos utilizados para o desenvolvimento deste trabalho, bem como as ferramentas utilizadas.

A metodologia adotada no estudo seguiu uma sequência estruturada de etapas para análise e comparação de dados de rastreamento ocular em diferentes condições (com e sem áudio). O processo inicia-se com a seleção da base de dados, que serviu como referência para as análises feitas. Em seguida, os dados foram pré-processados para as condições com áudio e sem áudio, o que garantiu que ambos os conjuntos fossem tratados de maneira adequada antes da análise. Após o pré-processamento, foram gerados histogramas das distribuições de fixação ocular, considerando as dimensões de longitude e latitude. Esses histogramas nos permitiram visualizar a concentração espacial das fixações em ambas as condições analisadas. Na etapa seguinte, foram calculados as médias e os desvios padrão de cada histograma, e a partir daí quantificamos a dispersão dos dados e realizamos uma descrição estatística mais precisa das distribuições de fixação. Com esses valores, realizamos os ajustes das curvas gaussianas para modelar a distribuição dos pontos de fixação em ambas as condições (com e sem áudio). A etapa seguinte consistiu na análise estatística entre as distribuições obtidas, onde identificamos diferenças significativas na distribuição das fixações entre as condições experimentais. Por fim, com base nos resultados dessa análise, discorremos sobre a comparação, onde extraímos as conclusões sobre o impacto da presença do áudio no comportamento visual dos participantes.

A Figura 2 ilustra o fluxo das fases da metodologia adotada nesta pesquisa, apresentando as etapas sequenciais que orientaram o desenvolvimento do estudo.

Figura 2: Fluxograma das etapas desenvolvidas neste estudo.



3.1 Base de Dados

Em nossos testes, usamos o conjunto de dados PAVS10K (Zhang, 2023). A escolha de tal conjunto de dados é justificada porque o sinal de áudio não é descartado durante seu procedimento de aquisição de dados, uma situação mais próxima de aplicações práticas da tecnologia imersiva.

O conjunto de dados inclui dados de rastreamento ocular de cerca de 20 observadores para cada um dos 67 vídeos disponíveis, que são agrupados em três categorias: speaking, music e miscellanea.

40 participantes (8 mulheres e 32 homens) com idades entre 18 e 34 anos, que relataram acuidade visual e auditiva normal ou corrigida para o normal. Vinte participantes foram selecionados aleatoriamente para assistir a vídeos com som mono (grupo 1), enquanto os demais assistiram a vídeos sem som (grupo 2). Os dois grupos possuem as mesmas distribuições de gênero e idade. Assim, cada vídeo com cada modalidade de áudio (ou seja, com ou sem som) foi visualizado por 20 participantes, e cada participante assistiu (livremente) cada vídeo apenas uma vez.

A Tabela 1 apresenta as principais características do conjunto usado, incluindo taxa de frames e número de frames.

Tabela 1: Características do subconjunto PAVS10K.

Video	*Rate (fps)*	Video	Rate (fps)
-0cfJOmUaNNI_1	30	-0cfJOmUaNNI_2	30
-0suxwissusc	30	-1An41lDIJ6Q	25
-1LM84FSzW0g_1	60	-1LM84FSzW0g_2	60
-1LM84FSzW0g_3	60	-4fxKBGthpaw	25
-4SilhsTuDU0	30	-5h95uTtPeck	24
-6QUCaLvQ_3I	60	-72f3ayGhMEA_1	50
-72f3ayGhMEA_2	50	-72f3ayGhMEA_3	50
-72f3ayGhMEA_4	50	-72f3ayGhMEA_6	50
_band	30	-bO43msZTfwA	30
-Bvu9m__ZX60	30	-ByBF08H-wDA	30
_conversation	30	_conversation2	30
-dBM3eM9HOoA	25	-eGGFGota5_A	30
-eqmjLZGZ36k	30	-ey9J7w98wlI	50
-ey9J7w98wlI_2	50	-g4fQ5iOVzsI	30
-G8pABGosD38	30	-Gq4Y4gL3zSg	24
-gTB1nfK-0Ac	30	-gy4TI-6j5po	30
-HNQMF7e6IL0	30	-I-43DzvhxX8_1	30
-idLVnagjl_s	30	-IRG9Z7Y2uS4	30
-J0Q4L68o3xE_1	30	-J0Q4L68o3xE_2	30
-jb5YxiXIsjU	25	-JBrp8aG4lro_1	30
-kZB3KMhqqyI	30	-LThm7VYvwxY	30
-MDwhMMSFkJ8_2	30	-MFVmxoXgeNQ	25
-MYmMZxmSc1U	25	-MzcdEI-tSUc_4	25
-n524y8uPUaU	30	-nDu57CGqbLM	30
-Ngj6C_RMK1g_1	30	-Ngj6C_RMK1g_2	30
-nZJGt3ZVg3g	30	-o7JEBWV4CmY	30
-Oak26yVbibQ	30	-P4KyjvsceZQ	30
-RbgxpagCY_c	30	-RrGhiInqXhc	30
-RSYbTSTz91g	30	_Ellen	30
_Ellen2	30	-SdGCX2H-_Uk	30
-SJIbpqgYWGw	30	-SZYXQ-6bfiQ	30
-TCUsegqBZ_M	25	-UpFZ2YaKeqM	30
-Uy5LTocHmoA	30	-V3zp7XOGBhs	30
ZuXCMpVR24I	30		

3.2 Procedimentos

O processamento e análise dos dados, a geração dos histogramas e o cálculo das médias e desvios padrão, foram conduzidos em Python, empregando bibliotecas como Pandas, OpenCV, NumPy e SciPy.

Para investigar o viés equatorial e o impacto do áudio na distribuição das fixações, foram utilizados testes estatísticos: test t student e teste F.

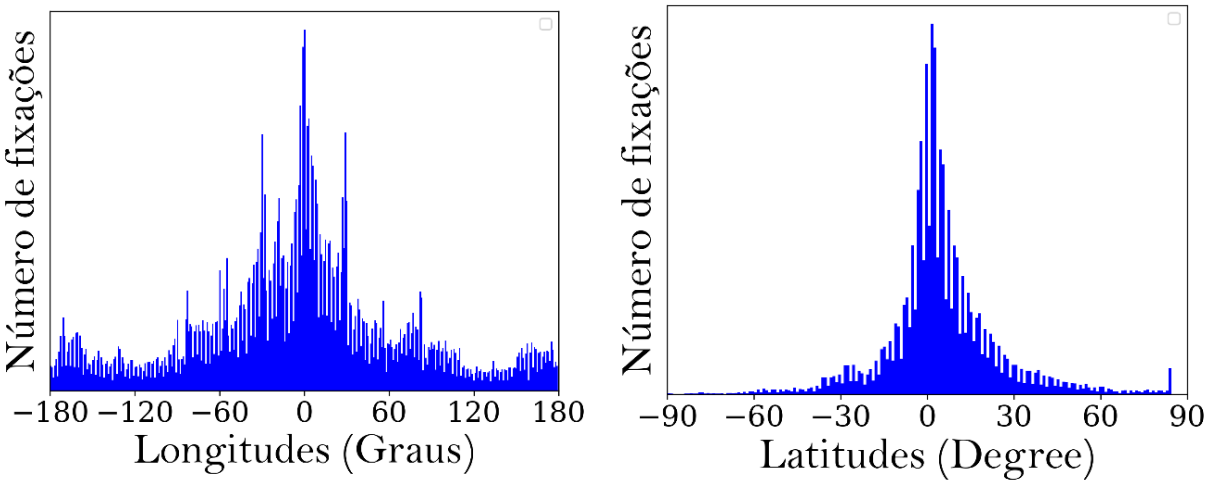
4. RESULTADOS E DISCUSSÕES

4.1 Viés Equatorial para dados com áudio

De acordo com Tatler (2007), o viés central pode ser encontrado independentemente de os elementos principais ou objetos salientes de uma cena serem centralmente enviesados. Então, para investigar se um viés central também pode ser observado no conjunto de dados PAVS10K (Zhang, 2023), analisamos os dados de fixação disponíveis.

Esta análise em particular se baseia em dois histogramas que mostram a distribuição de fixações nas direções longitudinal e latitudinal, abrangendo todo o conjunto de dados, ou seja, todos os sujeitos e estímulos. Construímos histogramas para direções longitudinais e latitudinais (Figura 3) a partir das fixações obtidas experimentalmente.

Figura 3: Histogramas de fixação geral com os dados de observadores que assistiram aos vídeos com áudio: (a) direção longitudinal; e (b) direção latitudinal.



Ao analisar o histograma longitudinal, percebe-se que há uma tendência dos observadores fixarem seu olhar de uma forma mais “equatorial". Da mesma forma, ao

inspecionar o histograma de fixações para a direção latitudinal, observamos novamente um aglomerado central, mas um decaimento mais acentuado, com uma maior concentração e distribuição na faixa de -20 a 20 graus.

A Tabela 2 apresenta a média e o desvio padrão avaliados para todos os vídeos de teste nas direções longitude e latitude.

Tabela 2: Média (μ) e desvio padrão (σ) para pontos de fixação com os dados de observadores que assistiram aos vídeos com áudio. Todos os espectadores, vídeos e frames são considerados.

	Longitude	**Latitude**
μ	-3.04	4.66
σ	76.85	19.81

As médias confirmam o Viés Equatorial, uma vez que seus valores permanecem relativamente próximos de zero 2,96 para longitude e 4,66 para latitude. Este último sugere uma tendência para valores um pouco mais altos que o Equador (no eixo Norte-Sul).

O desvio padrão, por sua vez, indica maior dispersão na direção longitudinal do que na direção latitudinal --- 76,85 para longitude e 19,81 para latitude. Este resultado confirma a existência de um Viés Equatorial, pois o maior desvio padrão na direção longitudinal sugere dispersão significativa de pontos de fixação ao longo de uma linha de visão logo acima do Equador; o valor médio da latitude sugere uma tendência a valores um pouco mais altos do que o Equador (no eixo Norte-Sul) e o maior desvio padrão na direção longitudinal sugere dispersão significativa de pontos de fixação ao longo do Equador.

A partir dos histogramas foi possível obter uma aproximação do viés equatorial e

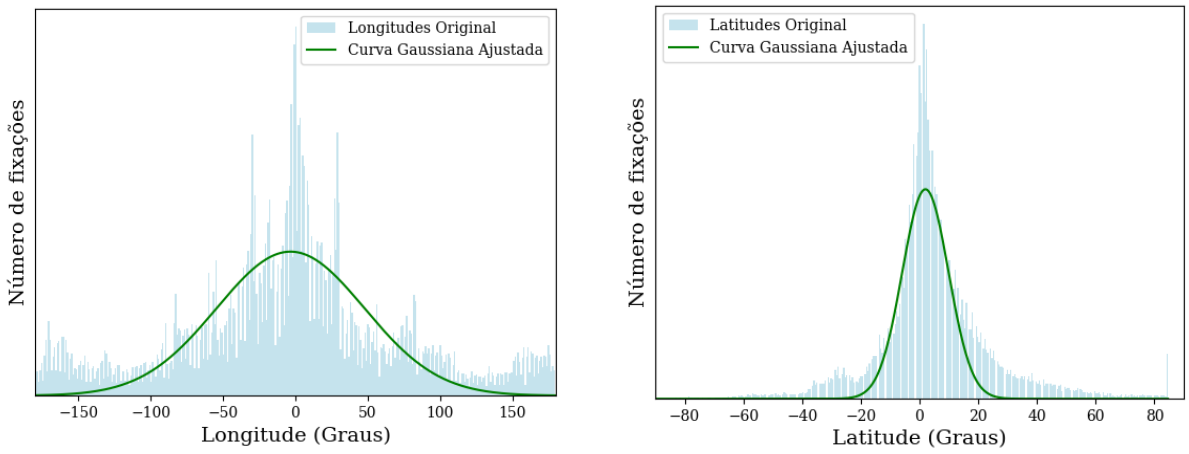
com os valores da médias e do desvio padrão, foi possível estimar o Viés Equatorial por funções Gaussianas 2D. A tabela 3 apresenta a média e o desvio padrão das curvas gaussianas ajustadas nas direções longitude e latitude.

Tabela 4: Média (μ) e desvio padrão (σ) para as curvas gaussianas ajustadas com os dados de observadores que assistiram aos vídeos com áudio.

	Longitude Ajustada	**Latitude Ajustada**
μ	-3.07	-1.53
σ	51.08	25.54

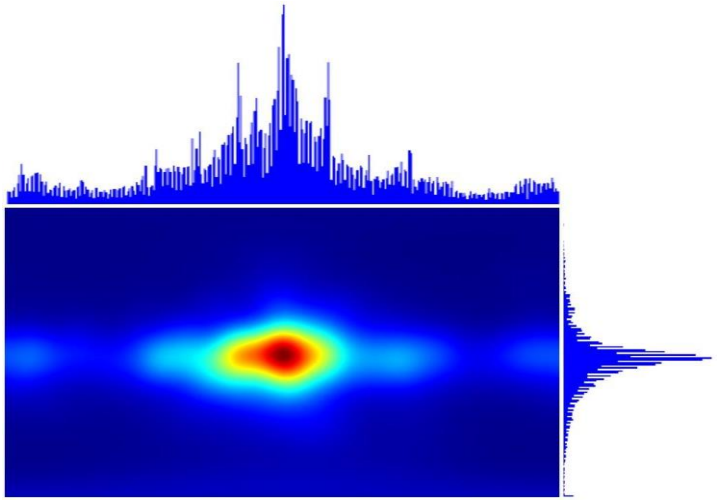
A Figura 4 ilustra as gaussianas geradas (em verde) a partir dos histogramas de fixações (em azul) nas direções de longitude e latitude.

Figura 4: Histograma (azul) e Gaussiano ajustado (verde). Histogramas gerados com os dados de observadores que assistiram aos vídeos com áudio.

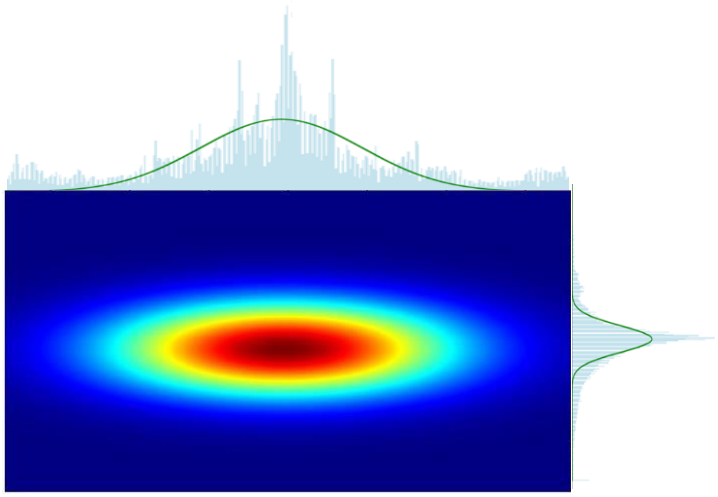


Enquanto a Figura 5 compara a aproximação do viés equatorial e o Viés Equatorial estimado no plano equirretangular.

Figura 5: (a) Aproximação de viés equatorial gerada a partir das fixações de todos os observadores que assitiram aos 67 vídeos com áudio. (b) Estimativa de viés equatorial gerada a partir do ajuste da gaussiana.



(a)



(b)

Outros trabalhos (Erwan, 2018 e Fang, 2020) realizaram essa análise e destacaram o uso de viés equatorial em modelos de saliência para vídeos 360. Conforme declarado por Xu et al (PATRICK, M., 2020), o viés equatorial pode ser usado como conhecimento prévio em modelos de atenção visual de vídeo/imagem 360° para melhorar a precisão da previsão. A existência de EB em vídeos imersivos é relevante, pois poderia

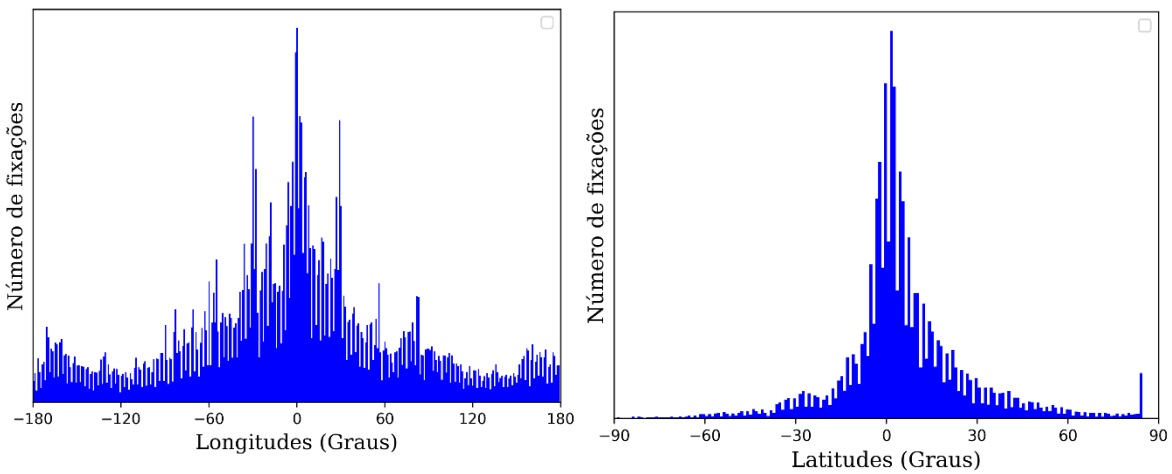
potencialmente melhorar os modelos de saliência ao incorporar esse conhecimento empírico em suas formulações ou, como pretendemos demonstrar, fornecer um referencial para a criação de novas abordagens que integrem o áudio como um fator determinante na distribuição das fixações oculares. Além disso, essa integração pode contribuir para o aprimoramento de técnicas de compressão de vídeos imersivos, priorizando regiões de maior interesse visual e otimizando a alocação de recursos computacionais. Essa abordagem, ao considerar a influência do áudio no direcionamento da atenção visual, pode representar um avanço significativo na modelagem de saliência e na experiência do usuário em ambientes de realidade virtual.

4.2 Viés Equatorial para dados sem áudio

Para investigar a presença do Viés Equatorial no conjunto de dados PAVS10K (Zhang, 2023) em condições sem áudio, procedemos na mesma maneira que em condições com áudio, analisando os dados de fixação disponíveis. Esta análise também se baseia em histogramas que mostram a distribuição de fixações nas direções longitudinal e latitudinal, abrangendo todos os sujeitos e estímulos.

A Figura 6 ilustra os histogramas para as direções longitudinal e latitudinal a partir das fixações.

Figura 6: Histogramas de fixação geral com os dados de observadores que assistiram aos vídeos sem áudio: (a) direção longitudinal; e (b) direção latitudinal.



Ao analisar o histograma longitudinal, observa-se que mesmo em condições sem áudio, os observadores tendem a fixar o olhar de forma mais "equatorial".

A Tabela 5 apresenta a média e o desvio padrão avaliados para todos os vídeos da base de dados, nas direções de longitude e latitude.

Tabela 5: Média (μ) e desvio padrão (σ) para pontos de fixação com dados de observadores que assistiram aos vídeos sem áudio. Todos os espectadores, vídeos e frames são considerados.

	Longitude	**Latitude**
μ	-1,36	5,67
σ	78,85	21,36

Assim como observado com na análise dos dados em condições com áudio, uma vez que seus valores permanecem relativamente próximos de zero: -1,36 para longitude e 5,67 para latitude, com desvio padrão maior na direção longitudinal do que na direção latitudinal — 78,85 para longitude e 21,36 para latitude.

Este resultado confirma a existência de um Viés Equatorial mesmo em condições sem áudio. Porém, podemos perceber que existe uma dispersão maior quando os observadores assistem aos vídeos sem o áudio, considerando o desvio padrão de ambos.

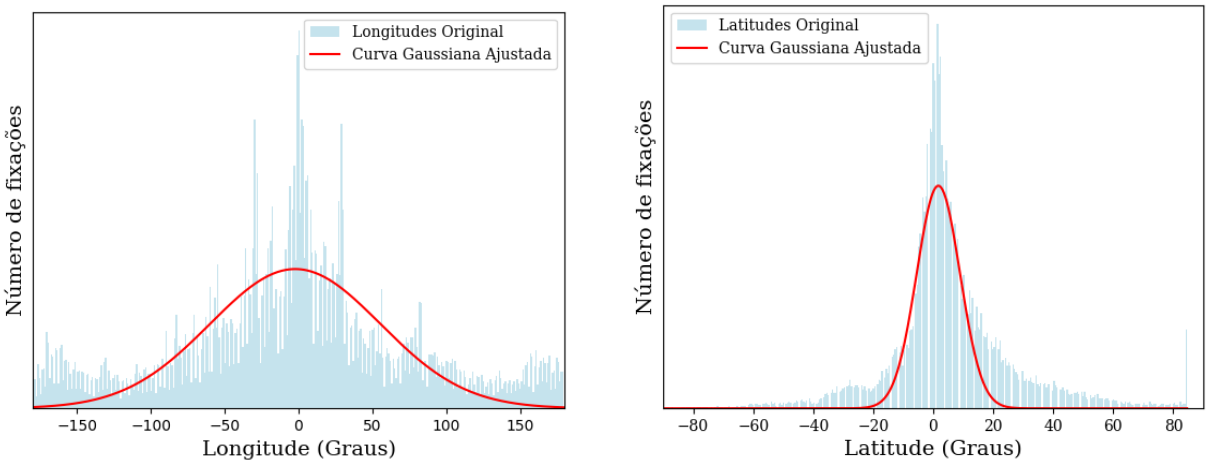
Ainda a partir dos histogramas, foi possível obter uma aproximação do Viés Equatorial e, com os valores das médias e do desvio padrão, estimar o Viés Equatorial por meio de funções Gaussianas 2D. A Tabela 6 apresenta a média e o desvio padrão das curvas gaussianas ajustadas nas direções longitude e latitude.

Tabela 6: Média (μ) e desvio padrão (σ) para as curvas gaussianas ajustadas com os dados de observadores que assistiram aos vídeos sem áudio.

	Longitude Ajustada	**Latitude Ajustada**
μ	-1.96	-0.98
σ	57.20	28.60

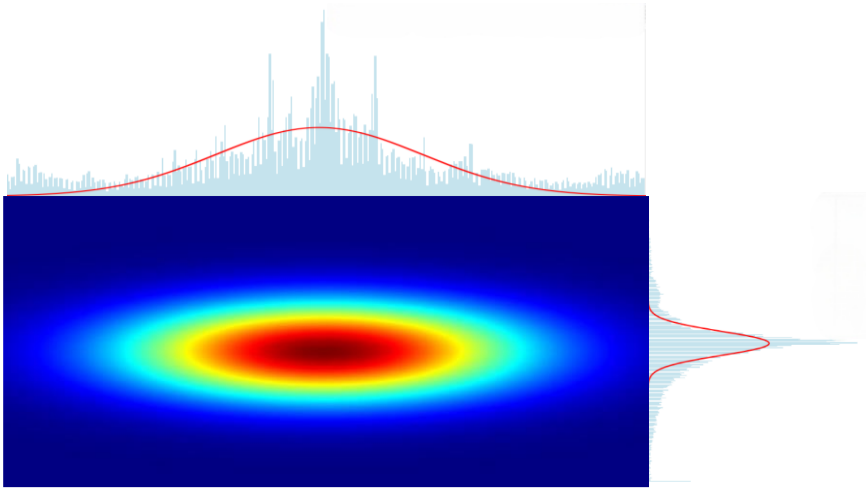
A Figura 7 ilustra as gaussianas geradas (em vermelho) a partir dos histogramas de fixações (em azul) nas direções de longitude e latitude.

Figura 7: Histograma (azul) e Gaussiana ajustada (vermelho). Histogramas gerados com os dados de observadores que assistiram aos vídeos sem áudio.



A Figura 8 ilustra o Viés Equatorial estimado a partir das gaussianas ajustadas com os dados dos observadores que assistiram aos vídeos sem aúdio.

Figura 8: Estimativa de viés equatorial gerada a partir do ajuste da gaussiana para os dados dos observadores que assistiram aos vídeos sem áudio.



4.3 Discussão e análise comparativa entre dados com e sem áudio

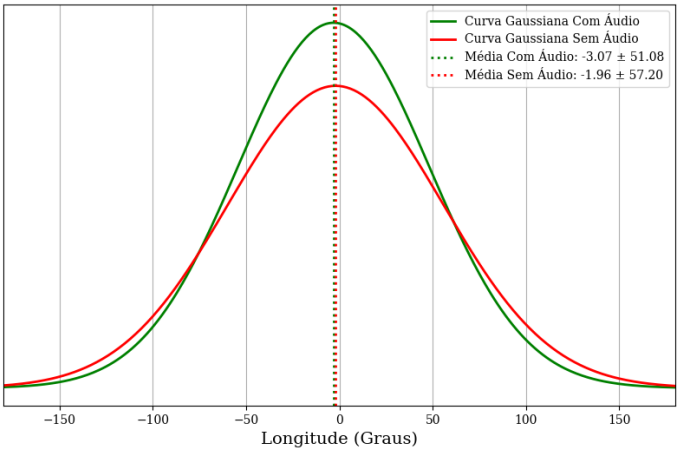
Ao comparar os resultados obtidos para os dados c om e sem áudio, observamos que as médias das fixações nas direções de longitude e latitude permanecem próximas de zero em ambos os casos, indicando uma tendência consistente dos observadores em fixar o olhar próximo ao equador da cena visual, mesmo em vídeos imersivos.

A Figura 9 (a) ilustra as duas curvas gaussianas ajustadas na longitude e a Figura 9 (b) as duas curvas gaussianas na latitude.

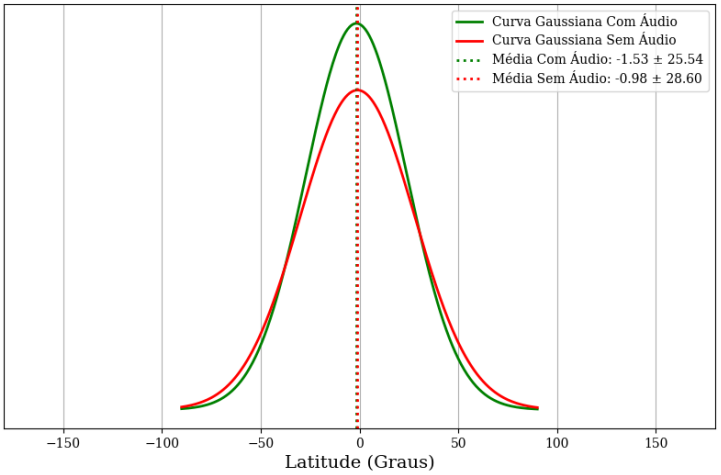
Como se pode perceber, existem variações nas médias, o que sugere que a presença ou ausência de áudio pode influencias a posição médias das fixações. Além disso, o desvio padrão na direção longitudinal é maior do que na direção latitudinal em ambos os cenários, indicando uma maior dispersão horizontal das fixações. Essa característica reforça a presença do Viés Equatorial, independentemente da condição auditiva.

Para determinar se as diferenças são estatisticamente significativas, realizamos testes estatísticos, como o teste t para as médias e o teste F para as variâncias. A Tabela 7 apresenta os valores para cada um dos testes.

Figura 9: Comparação entre as curvas gaussianas ajustadas.



(a) Curvas Gaussianas da longitude (graus).



(b) Curvas Gaussianas da latitude (graus)

Tabela 7: resultados para os Testes t-student e teste F.

	Longitude	**Latitude**
Teste t	t = -16.6630 p-valor = 2,4509e-62	t = -16.5129 p-valor = 2.9824e-61
Teste F	F = 0.7974 p-valor = 1.9999	F = 0.7974 p-valor = 1.9999

Através dos resultados obtidos, é possível perceber que tanto para a longitude quanto para a latitude, as médias das duas gaussianas são estatisticamente diferentes (p-valor próximo de zero). Isso indica que há um deslocamento sistemático nas distribuições (as médias das duas gaussianas não são iguais).

Já com relação as variâncias, não foram observadas diferenças significativas, ou seja, a dispersão dos dados em torno da média é semelhante entre as duas distribuições, tanto em longitude quanto em latitude (variâncias semelhantes).

As diferenças nas médias sugerem que as duas gaussianas representam distribuições com localizações distintas, ou seja, a presença ou ausência do áudio causa diferença significativa. Contudo, a ausência de diferença nas variâncias indica que a dispersão dos dados é semelhante entre as duas condições (com e sem áudio).

Os resultados obtidos foram analisados com foco na relação entre a presença do áudio e o viés equatorial em vídeos 360°. Observamos que a tendência de fixação predominante ao longo do eixo equatorial pode ser influenciada pelo áudio, que pode reforçar ou redistribuir o padrão de atenção visual. Uma menor dispersão das fixações e uma maior coerência grupal podem indicar que o áudio age como um guia atencional, direcionando o olhar para regiões específicas da cena.

Essas observações sugerem que o áudio exerce influência na distribuição das fixações, porém também podemos observar que o padrão geral do Viés Equatorial permanece consistente. Portanto, acreditamos que incorporar o Viés Equatorial como conhecimento prévio em modelos de saliência para vídeos 360° pode melhorar a precisão da previsão, independentemente da presença de áudio nos estímulos.

5. CONSIDERAÇÕES FINAIS

Este estudo buscou compreender o impacto do áudio na distribuição das fixações oculares em vídeos 360°, com foco na manifestação do viés equatorial. Os achados indicam que o áudio pode atuar como um fator modulador da atenção visual, influenciando a dispersão das fixações e sua concentração ao longo do eixo equatorial. Essa compreensão pode fornecer subsídios para o desenvolvimento de modelos computacionais de saliência mais robustos, além de contribuir para aplicações práticas, como otimização de conteúdo imersivo em realidade virtual, publicidade direcionada e treinamentos interativos.

Os resultados obtidos reforçam a importância de considerar estímulos multimodais na análise da atenção visual em cenários imersivos. Além disso, abrem caminho para pesquisas futuras que investiguem a interação entre áudio direcional e características visuais na modelagem da saliência em vídeos 360°.

Como proposta para trabalhos futuros, sugere-se uma análise mais detalhada das categorias de vídeos, verificando se a tendência equatorial se mantém em diferentes tipos de conteúdo. Essa abordagem pode fornecer uma compreensão mais aprofundada sobre a influência do contexto e das características específicas dos vídeos na distribuição das fixações oculares. A integração dessas descobertas pode aprimorar significativamente a experiência do usuário, permitindo o desenvolvimento de estratégias mais eficazes para guiar a atenção em ambientes virtuais.

REFERÊNCIAS BIBLIOGRÁFICAS

CHEN, M.; LI, Q.; ZHANG, S. Recent advances in 360° video streaming: A survey. *Multimedia Tools and Applications*, v. 77, n. 19, p. 24177-24203, 2018. DOI: 10.1007/s11042-018-6155-1.

CHENG, Y., TU, Z., MENG, F., ZHAI, J., & LIU, Y. (2018). Towards Robust Neural Machine Translation. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1756–1766.

DAVID, P. et al. Salient360! A dataset and methodology for salient object detection in 360° videos. *IEEE Transactions on Visualization and Computer Graphics*, v. 24, n. 12, p. 1-12, 2018. DOI: 10.1109/TVCG.2018.2794070.

DUCHOWSKI, Andrew T. *Eye tracking methodology: theory and practice*. 2. ed. Londres: Springer, 2007.

DRIVER, J. (2001). A selective review of selective attention research from the past century. *British Journal of Psychology*, 92(1), 53-78.

ERWAN, D. Equatorial bias in 360° video attention models. *Journal of Virtual Reality and Broadcasting*, v. 15, n. 3, p. 45-60, 2018.

FANG, Y. Saliency prediction in panoramic videos with equatorial bias. *Computer Vision and Pattern Recognition (CVPR)*, 2020. p. 987-996

FRATER, M. R., ARNOLD, J. F., & VAHEDIAN, A. (2001). Impact of Audio on Subjective Assessment of Video Quality in Videoconferencing Applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(9), 1059–1062. https://doi.org/10.1109/76.946522

LECUN, Y., BOTTOU, L., BENGIO, Y., & HAFFNER, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278-2324.

LI, Y. et al. A survey of visual attention models for 360° video. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Seoul: IEEE, 2019. p. 1-10.

NGUYEN, A.; YAN, Z.; NAHRSTEDT, K. (2018). "Your Attention is Unique: Detecting 360-Degree Video Saliency in Head-Mounted Display for Head Movement Prediction." *Proceedings of the 26th ACM International Conference on Multimedia (MM '18)*, pp. 1190–1198. https://doi.org/10.1145/3240508.3240678

PATRICK, M. (2020). Understanding attention mechanisms in neural networks. *Journal of Machine Learning Research*, 21(112), 1-35. Acesso em: 27 fev. 2025.

TATLER, Benjamin W. *Eye guidance in natural scenes: a new look at gaze*. In: LUXTON, David (Ed.). *Artificial vision: psychology and neurobiology*. Nova York: Oxford University Press, 2010. p. 9-32.

UPENIK, E.; ŘEŘÁBEK, M.; EBRAHIMI, T. (2016). "Testbed for Subjective Evaluation of Omnidirectional Visual Content." *2016 Picture Coding Symposium (PCS)*, pp. 1-5.

XU, Yanyu et al. Gaze prediction in dynamic 360° immersive videos. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. Disponível em: https://github.com/xuyanyu-shh/VR-EyeTracking. Acesso em: [data de acesso].

ZHANG, X. PAVS10K dataset. 2023. Disponível em: https://example-dataset-link.com. Acesso em: 27 fev. 2025.