

Modelagem de atenção audiovisual em vídeos imersivos: impactos do áudio na dispersão ocular

Felipe M. Ximenes, João Vitor R. Barros, Ronaldo F. Zampolo e Aline F. G. Sousa.

Resumo—Esta pesquisa investiga a influência do áudio na atenção visual em vídeos imersivos, através da análise da dispersão das fixações oculares. Utilizando dados de rastreamento ocular de cerca de 20 observadores, foram comparadas medidas de dispersão espacial das fixações oculares registradas para vídeos com e sem áudio. Os resultados indicam que o áudio reduz a variabilidade espacial das fixações, promovendo maior convergência da atenção visual. Esses achados reforçam a importância da integração multimodal na modelagem computacional da atenção visual e apontam o áudio como componente essencial no treinamento de modelos preditivos em vídeos imersivos.

Palavras-Chave—Modelagem de atenção audiovisual, Vídeos 360, Fixações.

Abstract—This research investigates the influence of audio on visual attention in immersive videos, through the analysis of the dispersion of eye fixations. Using eye tracking data from about 20 observers, measures of spatial dispersion of eye fixations recorded for videos with and without audio were compared. The results indicate that audio reduces the spatial variability of fixations, promoting greater convergence of visual attention. These findings reinforce the importance of multimodal integration in the computational modeling of visual attention and point to audio as an essential component in the training of predictive models in immersive videos.

Keywords—Audiovisual attention modeling, 360 videos, Fixations.

I. INTRODUÇÃO

Nas últimas décadas, muitos padrões de compressão de vídeo/imagem foram desenvolvidos para formatos 2D tradicionais. No entanto, devido às características esféricas dos vídeos imersivos, as abordagens 2D para armazenar, processar e transmitir vídeos mostraram-se subótimas para vídeo 360° [1]. 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 mídia [2], [3] para, assim, aliviar as cargas de armazenamento e transmissão, que exigem processamento dedicado e mais eficiente.

Com os avanços das pesquisas nesta área, alguns modelos de atenção visual foram propostos, porém a maioria deles usa apenas informações espaciais [4], [5], [6]. Sabe-se que a atenção não é atraída apenas por informações visuais, uma vez que vários outros sentidos influenciam na orientação da

atenção humana. A informação auditiva, por exemplo, é uma valiosa fonte de informação sensorial e sua influência na atenção visual foi comprovada em vários estudos, seja por experimentos psicológicos tradicionais seja por análises de rastreamento ocular [7].

O áudio, assim como a imagem, é um componente inerente aos vídeos. O impacto da informação de áudio na percepção de qualidade de vídeos 2D foi demonstrado por Frater, Arnold e Vahedian [7]. Já em vídeos imersivos, até onde se sabe, poucos modelos foram propostos levando em consideração elementos de áudio. Além disso, o interesse por estudos sobre modelos de atenção para vídeos 360° têm crescido por se verificar que conceitos e práticas consolidadas para mídias convencionais não têm o mesmo impacto em mídias imersivas.

Neste contexto, a presente pesquisa propõe avaliar o comportamento visual de observadores, através da análise da dispersão das fixações oculares em vídeos imersivos observados com e sem áudio.

II. DESENVOLVIMENTO

Apesar do avanço e da popularização dos vídeos imersivos, ainda há uma escassez de conjuntos de dados disponíveis para esse tipo de mídia. Alguns conjuntos de dados de predição de saliência visual foram propostos para imagens e vídeos imersivos [8], [9], [10], [2], mas nenhum deles considera elementos de áudio em suas análises e modelagem.

Na análise proposta nesta pesquisa, utiliza-se o conjunto de dados PAVS10K [11], que é composto por 67 vídeos imersivos. Todos os vídeos têm resolução 4K, formato equirretangular, taxa de frame variável e duração média de 29,6 segundos. Os vídeos podem ser agrupados nas seguintes categorias: *talking*, *music* e *miscellany*.

O PAVS10K contém dados de 40 participantes (8 mulheres e 32 homens), com idades entre 18 e 34 anos, que relataram acuidade visual e auditiva normal ou corrigida para normal. Vinte participantes foram selecionados aleatoriamente para assistir a vídeos com áudio mono, enquanto os demais assistiram a vídeos sem áudio. Cada participante assistiu livremente a cada vídeo apenas uma vez.

O conjunto de dados PAVS10K [11] foi utilizado por ser o conjunto de dados mais abrangente proposto até o momento. Foi necessário realizar pré-processamento dos dados e ajustá-los para realizar as análises. Na Figura 1 pode-se observar exemplos de alguns frames equirretangulares dos vídeos desta base de dados.

As análises foram realizadas de forma geral (com os 67 vídeos do conjunto de dados) e também para cada catego-

Felipe M. Ximenes; João Vitor R. Barros; Aline F. G. Sousa: Faculdade de Engenharia da Computação, Instituto de Geociências e Engenharias, Universidade Federal do Sul e Sudeste do Pará, Marabá-PA; e-mail: felipematteiximenes@gmail.com; joao956@unifesspa.edu.br; alinefarias@unifesspa.edu.br; Ronaldo F. Zampolo, Laboratório de Processamento de Sinais, Faculdade de Engenharia da Computação e Telecomunicações, Instituto de Tecnologia, Universidade Federal do Pará, Belém-PA, Brasil, email: zampolo@ufpa.br.

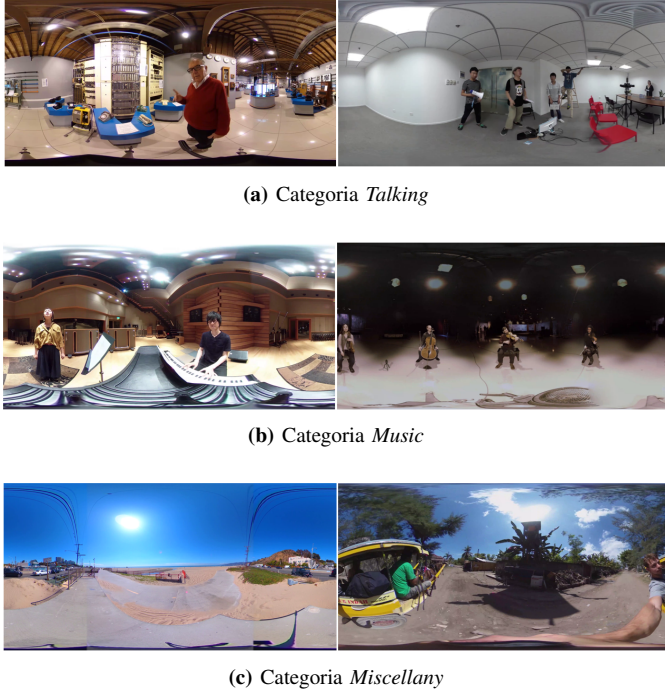


Fig. 1: Exemplos de frames equiretangulares por categoria.

ria: *talking* (35 vídeos), *music* (16 vídeos) e *miscellany* (16 vídeos).

A medida de dispersão das fixações foi calculada para cada frame dos vídeos. Como os vídeos possuem durações distintas, o tempo foi normalizado para uma escala comum de 0 a 1 e os dados foram interpolados utilizando interpolação linear, permitindo o cálculo das dispersões médias e a exibição dos resultados em uma escala temporal unificada.

O cálculo da dispersão das fixações em cada frame j é dado por:

$$D_j^m = \frac{1}{n_j^m} \sum_{i=1}^{n_j^m} (x_{ji}^m - \bar{x}_j^m)^2 + (y_{ji}^m - \bar{y}_j^m)^2 \quad (1)$$

onde $\bar{x}_j^m$ e $\bar{y}_j^m$ representam as coordenadas da fixação média no j -ésimo frame do m -ésimo vídeo; x_{ji}^m e y_{ji}^m denotam as coordenadas da i -ésima fixação no j -ésimo frame do m -ésimo vídeo; e n_j^m é o número total de fixações no j -ésimo frame do m -ésimo vídeo.

O cálculo da dispersão das fixações por categoria é dado por:

$$D_{\Phi_k}(j) = \frac{1}{M^k} \sum_{m \in \Phi_k} D_j^m, k \in \{geral, talking, music, misc\} \quad (2)$$

onde $D_{\Phi_k}(j)$ é a dispersão média das fixações oculares ao longo dos frames j para os vídeos pertencentes à categoria Φ_k ; e M^k representa o número de vídeos pertencentes à categoria Φ_k .

III. RESULTADOS NUMÉRICOS

A Figura 2 apresenta a dispersão ocular média, considerando todos os vídeos do *dataset*, ao longo do tempo normalizado

durante a visualização de vídeos imersivos com e sem a presença de áudio. Observa-se que, embora ambos os grupos apresentem aumento inicial da variância (associado ao início da exploração visual), os participantes expostos ao áudio (linha laranja) mantêm uma menor dispersão em comparação com aqueles na condição silenciosa (linha azul tracejada).

Esse comportamento sugere que a presença do áudio atua como um fator de guia atencional, direcionando os olhares para regiões mais específicas da cena e, portanto, reduzindo a variabilidade espacial das fixações oculares ao longo do tempo.

Além disso, foram realizadas análises também por categorias, para avaliar se a presença do áudio mantém a mesma tendência geral. Os resultados alcançados nas três categorias (Figuras 2 (b)-(d)) confirmam o áudio como um elemento redutor da dispersão das fixações oculares independente do conteúdo dos vídeos.

A Tabela I apresenta a média temporal dos gráficos apresentados na Figura 2. Observa-se uma redução da dispersão ocular na condição com áudio em todas as categorias analisadas.

TABELA I: Taxa de dispersão média normalizada por categoria, com e sem áudio.

Categoria	Com áudio	Sem áudio
<i>Talking</i>	4120,62	5008,98
<i>Music</i>	4728,06	5946,16
<i>Miscellany</i>	5862,86	6374,93
Média Geral	4681,74	5558,99

A categoria *Talking* apresentou a maior diferença absoluta, com uma redução de aproximadamente 17,70% na variância média quando o áudio está presente (4120,62 com áudio vs. 5008,98 sem áudio). Esse resultado sugere que a presença de fala atua como um forte direcionador atencional, promovendo a convergência do olhar para regiões específicas do campo visual, como rostos ou bocas em movimento.

Na categoria *Music*, a diferença entre as condições foi a maior observada, com uma redução da dispersão em 20,48% na presença de áudio (4728,06 com áudio vs. 5946,16 sem áudio). Essa discrepância pode estar relacionada ao fato de que vídeos musicais, que envolvem estímulos auditivos contínuos, apresentam correspondência direta com eventos visuais localizados.

Já na categoria *Miscellany*, composta por vídeos diversos com menor estrutura narrativa ou foco visual centralizado, foi observada uma diferença menos significativa, com uma redução de cerca de 8,08% na dispersão média com o uso do áudio. Ainda assim, acredita-se que esse comportamento indica que, mesmo em contextos mais diversos, o áudio oferece pistas relevantes que guiam a atenção visual.

De forma generalizada, o cálculo da média geral apresentado da tabela, evidencia uma redução média de aproximadamente 15,78% na dispersão ocular quando o áudio está presente (4681,74 com áudio vs. 5558,99 sem áudio), reforçando a hipótese de que informações auditivas contribuem para modular e concentrar a atenção visual em vídeos imersivos. Além disso, foi aplicado o teste t de Student para comparar a dispersão ocular média entre os grupos com e sem áudio. Os resultados indicaram uma diferença estatisticamente significativa com $t = -5,16$ e $p = 7,9510^{-6}$ ($p < 0.05$).

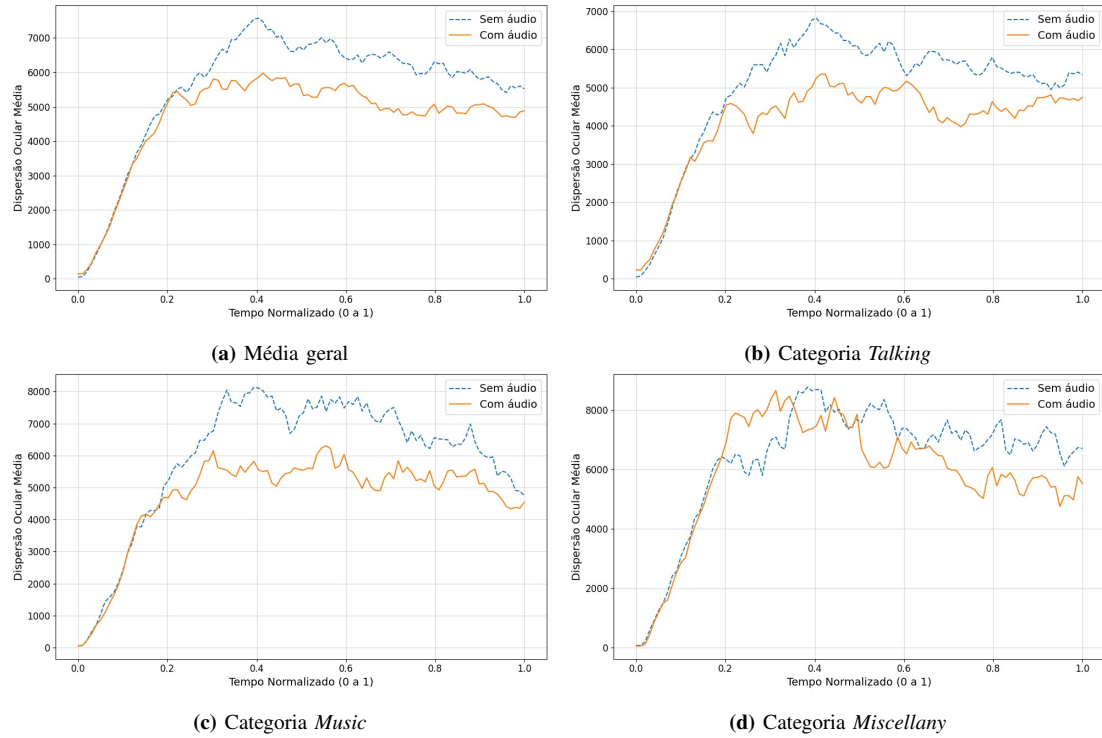


Fig. 2: Curvas das dispersões oculares.

Esses resultados estão alinhados com a literatura para vídeos 2D que aponta o papel do áudio como um guia atencional relevante em ambientes multimodais, independente da categoria.

IV. CONCLUSÕES

Este artigo analisou o comportamento da atenção visual em vídeos imersivos com base em dados de rastreamento ocular, comparando a dispersão das fixações em cenários com e sem áudio. A presença de estímulos sonoros demonstrou reduzir essa dispersão, indicando maior convergência da atenção visual em vídeos acompanhados de som.

Esses resultados sugerem que o áudio atua como direcionador atencional, funcionando como um ajuste modulador (um "offset") que estreita o foco dos observadores e os orienta para regiões específicas da cena. Tal efeito evidencia a importância da integração de informações auditivas na modelagem computacional da atenção em ambientes imersivos.

A influência do áudio foi especialmente importante em vídeos com conteúdo musical, que envolve estímulos auditivos contínuos, revelando uma forte conexão entre os estímulos auditivos e visuais no comportamento ocular. Essa relação reforça a necessidade de considerar aspectos multimodais na modelagem da atenção visual para realidade virtual e vídeos 360.

Como direções futuras, recomenda-se o desenvolvimento de modelos computacionais que integrem informações auditivas e visuais, permitindo uma representação mais realista da atenção humana em ambientes multimodais.

REFERÊNCIAS

- [1] Z. Chen, Y. Li, and Y. Zhang, "Recent advances in omnidirectional video coding for virtual reality: Projection and evaluation," *Signal Processing*, vol. 146, pp. 66–78, 2018.
- [2] Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao, "Gaze prediction in dynamic 360 immersive videos," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5333–5342.
- [3] Mai Xu, Chen Li, Shanyi Zhang, and Patrick Le Callet, "State-of-the-art in 360° video/image processing: Perception, assessment and compression," *IEEE Journal of Selected Topics in Signal Processing*, vol. PP, pp. 1–1, 01 2020.
- [4] Hsien-Tzu Cheng, Chun-Hung Chao, Jin-Dong Dong, Hao-Kai Wen, Tyng-Luh Liu, and Min Sun, "Cube padding for weakly-supervised saliency prediction in 360° videos," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 1420–1429.
- [5] Anh Nguyen, Zhisheng Yan, and Klara Nahrstedt, "Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction," in *Proceedings of the 26th ACM International Conference on Multimedia*, New York, NY, USA, 2018, MM '18, p. 1190–1198, Association for Computing Machinery.
- [6] Minglang Qiao, Mai Xu, Zulin Wang, and Ali Borji, "Viewport-dependent saliency prediction in 360° video," *IEEE Transactions on Multimedia*, vol. 23, pp. 748–760, 2021.
- [7] Michael R. Frater, John F. Arnold, and Abedin Vahedian, "Impact of audio on subjective assessment of video quality in videoconferencing applications," *IEEE transactions on circuits and systems for video technology*, vol. 11, no. 9, pp. 1059–1062, 2001.
- [8] Anh Nguyen and Zhisheng Yan, "A saliency dataset for 360-degree videos," in *Proceedings of the 10th ACM Multimedia Systems Conference*, New York, NY, USA, 2019, MMSys '19, p. 279–284, Association for Computing Machinery.
- [9] Xavier Corbillon, Francesca De Simone, and Gwendal Simon, "360-degree video head movement dataset," New York, NY, USA, 2017, MMSys'17, p. 199–204, Association for Computing Machinery.
- [10] Yashas Rai, Jesús Gutiérrez, and Patrick Le Callet, "A dataset of head and eye movements for 360 degree images," in *Proceedings of the 8th ACM on Multimedia Systems Conference*, New York, NY, USA, 2017, MMSys'17, p. 205–210, Association for Computing Machinery.
- [11] Yi Zhang, Fang-Yi Chao, Wassim Hamidouche, and Olivier Deforges, "PAV-SOD: A new task towards panoramic audiovisual saliency detection," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 3, pp. 1–26, 2023.