

Análise baseada em IoU de *viewports* em vídeos imersivos: implicações do áudio para a modelagem de saliência visual

Felipe C. Carvalho, Ronaldo F. Zampolo e Aline F. G. Sousa

Resumo—O desenvolvimento de modelos de atenção visual para ambientes imersivos é essencial para sistemas baseados em Realidade Virtual (RV). Neste trabalho, analisamos a convergência das *viewports* em vídeos 360° por meio da *Intersection over Union* (IoU), utilizada para quantificar a sobreposição entre regiões visualizadas por diferentes observadores. Com dados de rastreamento ocular da base PAVS10K (67 vídeos), avaliamos o efeito do áudio na atenção visual. Os resultados indicam que a presença de áudio aumenta em cerca de 18% a sobreposição entre *viewports*. Até onde sabemos, este é o primeiro estudo a empregar a IoU nessa análise em vídeos omnidirecionais. Esses achados reforçam a relevância de integrar informações de áudio na modelagem de saliência visual em vídeos imersivos.

Palavras-Chave—Vídeos 360°, modelagem de saliência audio-visual, *viewports*, realidade virtual, atenção audiovisual.

Abstract— *The development of visual attention models for immersive environments is essential for virtual reality-based systems. In this work, we analyze the convergence of viewports in 360° videos using the Intersection over Union (IoU) metric to quantify the overlap between regions viewed by different observers. Using eye-tracking data from the PAVS10K database (67 videos), we evaluate the effect of audio on visual attention. The results indicate that audio increases viewport overlap by approximately 18%. To the best of our knowledge, this is the first study to employ IoU for this analysis in omnidirectional videos. These findings reinforce the relevance of integrating audio information into visual saliency modeling for immersive videos.*

Keywords—360° videos, audiovisual saliency modeling, *viewports*, virtual reality, audiovisual attention.

I. INTRODUÇÃO

A popularização de sistemas baseados em Realidade Virtual (RV), aliada ao amplo acesso a dispositivos de captura de vídeos em 360°, tem impulsionado a criação e o consumo de conteúdos imersivos. Diferentemente dos vídeos tradicionais, onde o enquadramento define o que será visto, em ambientes omnidirecionais o observador é livre para selecionar sua própria janela de visualização, comumente chamada de *viewport*.

Essa característica introduz novos desafios para a compreensão da atenção visual em ambientes imersivos. Enquanto em mídias convencionais a análise da atenção pode ser realizada diretamente sobre a região exibida na tela, em vídeos 360°

é necessário considerar o comportamento exploratório dos observadores. Nesse contexto, a modelagem de saliência visual surge como uma ferramenta fundamental para prever regiões relevantes e compreender a seleção das *viewports*.

Historicamente, a modelagem de saliência visual teve sua gênese focada em mídias 2D tradicionais. Modelos pioneiros baseavam-se primariamente na Hipótese de Integração de Características [1], extraindo atributos de baixo nível das imagens para prever as fixações oculares.

Com o advento dos vídeos 360° e da RV, a liberdade de exploração dos observadores, aliada às descontinuidades espaciais e distorções geométricas da projeção equirretangular [2], passou a representar um novo desafio para a modelagem da atenção visual.

Outro aspecto fundamental é a influência de múltiplas modalidades sensoriais. Embora estudos em psicologia cognitiva já tenham demonstrado qualitativamente que estímulos auditivos podem modificar a percepção e a experiência do observador [3], ainda existem lacunas quanto à quantificação desse efeito em vídeos 360°. Além disso, grande parte das avaliações existentes ainda herda conceitos e técnicas originalmente desenvolvidos para o domínio 2D.

Diante dessa lacuna, este trabalho propõe uma análise comparativa de *viewports* em vídeos 360°, utilizando a métrica *Intersection over Union* (IoU) para quantificar a sobreposição entre regiões observadas. Especificamente, são comparados grupos de observadores submetidos a vídeos com e sem áudio, buscando investigar de que forma o componente sonoro influencia a convergência da atenção visual.

Embora a IoU seja amplamente utilizada em tarefas de visão computacional, incluindo segmentação semântica e detecção de objetos [6], a revisão da literatura não identificou trabalhos que a empreguem para quantificar diretamente a sobreposição entre *viewports* de múltiplos observadores em vídeos omnidirecionais. Até onde é de nosso conhecimento, esta constitui a primeira investigação que utiliza a IoU para esse propósito.

As principais contribuições deste trabalho incluem:

- a utilização da métrica IoU para quantificar a sobreposição de *viewports* em projeções equirretangulares, lidando com descontinuidades espaciais;
- a análise quantitativa do impacto de estímulos sonoros na convergência da atenção visual em diferentes categorias de vídeos imersivos.

II. METODOLOGIA

A. Mapeamento e determinação de viewports

Foi utilizado o conjunto de dados PAVS10K [4], composto por 67 vídeos omnidirecionais acompanhados de sinais de áudio e registros de rastreamento ocular. A escolha dessa base deve-se à sua metodologia de aquisição, que preserva o sinal de áudio original dos vídeos. Os dados foram coletados de 40 participantes, divididos em dois grupos de 20 indivíduos. No grupo 1, os vídeos foram apresentados com áudio, enquanto no grupo 2 foram exibidos sem áudio. Cada participante assistiu a cada vídeo apenas uma vez, em regime de visualização livre. Os vídeos estão organizados em três categorias: *speaking* (35 vídeos), *music* (16 vídeos) e *miscellanea* (16 vídeos).

Os vídeos são representados na projeção equiretangular, que mapeia a esfera em um plano bidimensional. Neste trabalho, a *viewport* é definida como um retângulo delimitador projetado sobre esse plano, com campo de visão (*Field of View* – FoV) de 120° na horizontal e 90° na vertical. Essa configuração aproxima a região de sobreposição da visão binocular humana e é compatível com limitações físicas de *displays* de RV comerciais [5]. A Figura 1 ilustra a projeção de duas *viewports* em um frame equiretangular.

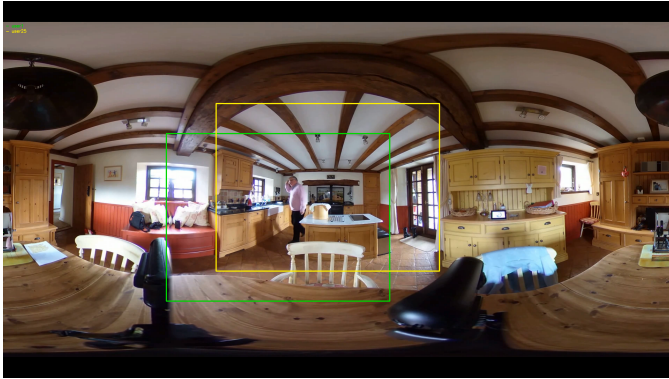


Fig. 1. Exemplo da projeção equiretangular ilustrando as *viewports* de dois observadores (identificados por retângulos coloridos).

B. Cálculo de IoU entre viewports

O cálculo da IoU entre *viewports* foi realizado sobre a projeção equiretangular. Para lidar com a descontinuidade geométrica introduzida por esta projeção, antes do cálculo da IoU, foi implementado um algoritmo de translação espacial das *viewports* que atua tanto nos limites laterais ($\pm 180^\circ$) quanto nos limites verticais ($\pm 90^\circ$). Quando as coordenadas de projeção de uma *viewport* ultrapassam qualquer um desses limites, o algoritmo gera dinamicamente réplicas geométricas, transladadas em $\pm 360^\circ$ (horizontalmente) e $\pm 180^\circ$ (verticalmente). Essa estratégia preserva a continuidade espacial da esfera, garantindo a intersecção correta entre regiões que, embora separadas na projeção plana, são adjacentes no domínio esférico, conforme ilustrado na Figura 2.

Com isso, a formulação matemática da IoU (Equação 1) permanece inalterada, sendo modificada apenas a representação geométrica das *viewports* para preservar a continuidade da superfície esférica. Neste trabalho, utiliza-se a notação IoU_{esf}



Fig. 2. Representação das réplicas transladadas de uma *viewport* para correção das descontinuidades da projeção equiretangular no domínio esférico.

apenas para indicar que a IoU é calculada sobre *viewports* previamente corrigidas por esse procedimento. Dadas as *viewports* de dois observadores distintos, representadas pelas áreas geométricas A e B , a IoU_{esf} é definida pela razão entre a área de sobreposição de ambas e a área total combinada:

$$IoU_{esf}(A, B) = \frac{\text{Área}(A \cap B)}{\text{Área}(A \cup B)}, \quad (1)$$

Para quantificar a convergência do grupo em um determinado instante (frame t), calcula-se a média aritmética do IoU_{esf} de todos os pares de observadores naquele exato momento. Dado um conjunto de N observadores, essa média é calculada como:

$$\overline{IoU}_{esf}(t) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^N IoU_{esf}(V_i(t), V_j(t)), \quad (2)$$

onde N representa o número total de observadores no grupo avaliado, enquanto $V_i(t)$ e $V_j(t)$ correspondem às *viewports* espaciais projetadas dos observadores i e j , respectivamente, no frame t . Dessa forma, $\overline{IoU}_{esf}(t)$ representa o grau médio de convergência entre todos os pares de observadores no frame t , fornecendo uma medida da atenção visual coletiva do grupo ao longo do tempo.

III. RESULTADOS E DISCUSSÃO

Inicialmente, para compreender as nuances comportamentais em diferentes tipos de conteúdo, foi realizada a análise nas três categorias: *speaking*, *music* e *miscellanea*.

A Tabela I resume os valores médios de $\overline{IoU}_{esf}$ para cada categoria de vídeo nas condições com e sem áudio, além do aumento percentual proporcionado pelo áudio e dos resultados do teste de Wilcoxon pareado, empregados para verificar se as diferenças entre as condições são estatisticamente significativas.

A métrica IoU_{esf} revelou uma tendência consistente de maior sobreposição entre as *viewports* dos observadores que assistiram aos vídeos com áudio, o que indica uma maior convergência da atenção visual na presença de estímulos sonoros. Em termos globais, a introdução do áudio elevou

TABELA I

COMPARAÇÃO DOS VALORES MÉDIOS DE $\overline{IoU}_{esf}$ ENTRE AS CONDIÇÕES COM E SEM ÁUDIO PARA CADA CATEGORIA DE VÍDEO, INCLUINDO O GANHO PERCENTUAL E O RESULTADO DO TESTE DE WILCOXON PAREADO.

Categoria	Sem áudio ($\mu \pm \sigma$)	Com áudio ($\mu \pm \sigma$)	Aumento	p-valor
<i>Speaking</i> (35)	0,407 $\pm$ 0,080	0,490 $\pm$ 0,096	+20,47%	< 0,0001
<i>Music</i> (16)	0,356 $\pm$ 0,056	0,420 $\pm$ 0,072	+18,01%	0,0004
<i>Miscellanea</i> (16)	0,388 $\pm$ 0,050	0,435 $\pm$ 0,082	+12,09%	0,0021
Global (67)	0,390 $\pm$ 0,071	0,460 $\pm$ 0,092	+17,94%	< 0,0001

a sobreposição média de 0,390 para 0,460, correspondendo a um aumento de 17,94%. O teste de Wilcoxon pareado confirmou que esse efeito é estatisticamente significativo em todas as categorias ($p < 0,001$ para o conjunto global e para *speaking*; $p < 0,01$ para *music* e *miscellanea*), fornecendo evidências para rejeitar a hipótese nula de igualdade entre as condições com e sem áudio.

A categoria *speaking* apresentou o maior ganho (+20,47%), seguida por *music* (+18,01%), enquanto *miscellanea* registrou o menor impacto (+12,09%).

Para ilustrar os extremos do comportamento observado, a Tabela II apresenta os valores médios de IoU para seis vídeos selecionados (dois de cada categoria), correspondentes aos casos de maior e menor sobreposição de *viewports*.

TABELA II

VALORES MÉDIOS DE IoU PARA SEIS VÍDEOS REPRESENTATIVOS DENTRE OS 67 ANALISADOS.

Categoria	Vídeo	Sem áudio	Com áudio	Aumento
<i>Speaking</i>	—72f3ayGhMEA_2	0,3803	0,6059	+59,31%
	—n524y8uPUaU	0,5064	0,4486	-11,40%
<i>Music</i>	—sdGCX2H—_Uk	0,3294	0,5009	+52,05%
	—gTB1nfK—0Ac	0,3690	0,3308	-10,35%
<i>Miscellanea</i>	—Ngj6C_RMK1g_2	0,3196	0,4309	+34,81%
	—P4KyjvsceZQ	0,3850	0,3398	-11,75%

Embora os resultados globais indiquem que a presença do áudio aumenta a convergência das *viewports*, os casos individuais apresentados na Tabela II revelam que esse comportamento não é universal. Em alguns vídeos, o áudio reduziu a sobreposição entre as *viewports* observadas. Esse fenômeno sugere que a influência sonora sobre a atenção visual depende da natureza do estímulo acústico e de sua relação com os elementos visuais da cena. Quando o áudio possui uma associação espacial clara, ele tende a atuar como vetor de convergência atencional entre observadores. Entretanto, estímulos sonoros mais distribuídos podem estimular diferentes estratégias exploratórias, aumentando a dispersão das *viewports*. Dessa forma, o áudio pode ser interpretado como um modulador do comportamento exploratório em ambientes imersivos.

Do ponto de vista perceptual, seres humanos são naturalmente predispostos a buscar a fonte visual de estímulos sonoros, especialmente quando associados à fala [7]. Na categoria *speaking*, a presença de um locutor atua como um vetor de atração, alinhando rapidamente o foco dos observadores e reduzindo a exploração periférica.

Na categoria *music*, o efeito é moderado, já que o palco sonoro tende a ser mais distribuído espacialmente, diluindo o direcionamento da atenção. Por fim, na categoria *miscellanea*,

a ausência de uma fonte sonora bem definida resulta em menor influência do áudio na organização da atenção visual.

IV. CONCLUSÕES

Este trabalho apresentou uma abordagem inédita para a análise comparativa de *viewports* em vídeos imersivos, empregando a métrica IoU para mensurar o grau de sobreposição entre as regiões observadas pelos usuários. Os resultados demonstraram que a comparação entre *viewports* permite identificar padrões de convergência da atenção visual que não podem ser capturados por abordagens tradicionais herdadas de vídeos bidimensionais, evidenciando a necessidade de metodologias específicas para ambientes 360°.

A análise de 67 vídeos evidenciou que a presença de estímulos sonoros eleva, em média, em aproximadamente 18% essa convergência, indicando o papel do áudio como mecanismo de direcionamento atencional. Além disso, a análise por categoria revelou que esse efeito não é uniforme, sendo influenciado pelo conteúdo e pela natureza da fonte sonora.

Os achados deste estudo ampliam a compreensão sobre o comportamento visual em vídeos imersivos e oferecem subsídios para o desenvolvimento de modelos computacionais de saliência, bem como para aplicações em *streaming* adaptativo, avaliação de Qualidade de Experiência (QoE) e criação de conteúdos para RV.

Como trabalhos futuros, destacam-se a investigação do impacto de diferentes parametrizações da FoV e de diferentes condições do áudio no direcionamento da atenção visual. Além disso, como perspectiva futura, pretende-se investigar métricas de sobreposição definidas diretamente na esfera, de forma a minimizar as distorções introduzidas pela projeção planar.

AGRADECIMENTOS

Este estudo tem o suporte financeiro da Fundação Amazônia de Amparo a Estudos e Pesquisas (FAPESPA). Agradecimentos ao grupo de pesquisa em Visão Computacional e Aprendizado de Máquina (VICAM) da Universidade Federal do Sul e Sudeste do Pará (UNIFESSPA) pelo suporte técnico.

REFERÊNCIAS

- [1] L. Itti, C. Koch e E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
- [2] A. Nguyen, Z. Yan e K. Nahrstedt, "Your Attention is Unique: Predicting 360-Degree Video Saliency in Head-Mounted Display," em *Proceedings of the 26th ACM International Conference on Multimedia*, pp. 1190–1198, 2018.
- [3] P. Larsson, D. Västfjäll e M. Kleiner, "Effects of auditory information on presence in virtual environments," *Presence: Teleoperators and Virtual Environments*, vol. 16, no. 1, pp. 63–66, 2007.
- [4] Y. Zhang, F.-Y. Chao, W. Hamidouche e O. Déforges, "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.
- [5] S. M. LaValle, *Virtual Reality*, 1st ed. Cambridge, U.K.: Cambridge University Press, 2017.
- [6] M. Everingham, L. Van Gool, C. K. Williams, J. Winn e A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [7] E. C. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953.