Mais de um sentido: Facilitando a autoria, sincronização e execução de efeitos sensoriais em linguagens multimídia

Raphael Abreu
raphael.abreu@midiacom.uff.br
CEFET/RJ
Laboratório Midiacom - UFF

Eduardo Bezerra
ebezerra@cefet-rj.br
CEFET/RJ

Joel Santos
jsantos@eic.cefet-rj.br
CEFET/RJ

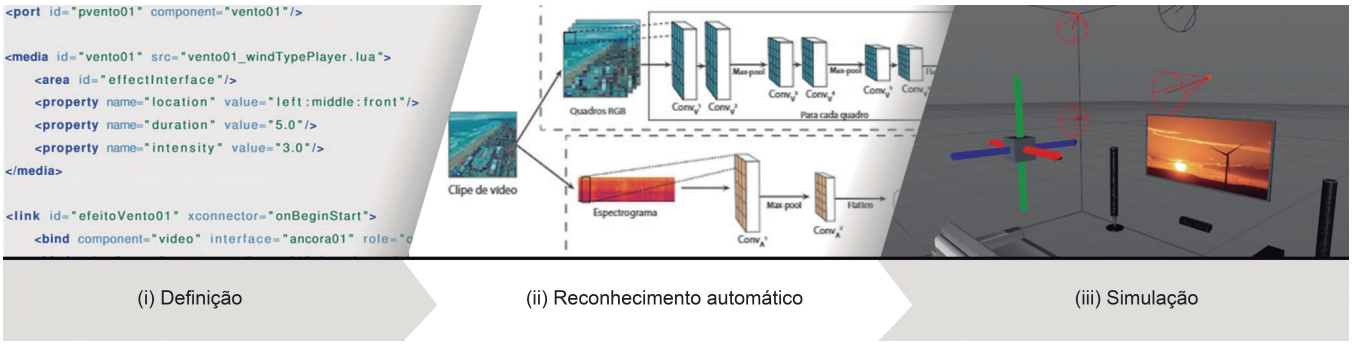

Figure 1: Principais contribuições da dissertação. Inicialmente, uma forma de definição de efeitos sensoriais em NCL, em conjunto com uma definição genérica da sincronização dos mesmos. Em seguida, uma rede neural bimodal construída e treinada para reconhecimento de efeitos sensoriais. Por fim um simulador em 3D que permite visualização da ativação de efeitos sensoriais.

ABSTRACT

Advances in ubiquitous computing have spurred research aimed at increasing user immersion in multimedia applications. One such research is the coupling of sensory effects in multimedia presentations. That is, to enable forms of interaction with human senses other than usual sight and hearing. However, the development of applications with sensory effects needs an authoring effort to synchronize sensory effects with audiovisual content. In addition, interactive applications lack abstractions to facilitate the authoring of sensory effects. Therefore, this paper presents works on three approaches to facilitate this scenario.

KEYWORDS

mulsemedia, NCL, neural networks, simulation, 3D

1 INTRODUÇÃO

Este artigo apresenta um resumo dos resultados obtidos na dissertação defendida no Programa de Pós-graduação em julho de 2018. O trabalho foi desenvolvido pelo 1º autor, num período de 22 meses, sob a orientação dos últimos autores. O trabalho desenvolvido na dissertação está relacionado as áreas da computação de ambientes virtuais, sistemas multimídia, *mulsemedia* e aprendizado de máquina. O trabalho desenvolvido foca em propor abstrações e ferramentas que facilitem a autoria, sincronização e execução de efeitos sensoriais em linguagens multimídia.

In: I Concurso de Teses e Dissertações (CTD 2019), Rio de Janeiro, Brasil. Anais Estendidos do Simpósio Brasileiro de Sistemas Multimidia e Web (WebMedia). Porto Alegre: Sociedade Brasileira de Computação, 2019.
ISSN 2596-1683

O trabalho se mostra necessário e atual pois facilita a definição de aplicações *mulsmedia* interativas. Aplicações *mulsmedia* realizam interface com outros sentidos humanos além da visão e audição [2], que por sua vez têm grande potencial de aumentar a imersão do usuário no conteúdo apresentado. Um conteúdo que atinge outros sentidos humanos diferentes de visão e audição é geralmente chamado de efeito sensorial. Por exemplo, um aroma ou um vento. Porém aplicações contendo efeitos sensoriais requerem um grande esforço de autoria para realizar a sincronização entre eles e o conteúdo audiovisual. Este esforço está relacionado seja à indicação de quando um efeito deverá ser executado [9], quanto a caracterização de um efeito na aplicação [8].

Para indicar quando um efeito sensorial deve ser executado, o autor deve sincronizar momentos de renderização de efeitos sensoriais com momentos de apresentação de "componentes de cena" em uma mídia. Um componente de cena pode ser um objeto ou um conceito que esteja sendo representado na mídia. Para sincronizar um efeito sensorial, por exemplo, com o componente de cena praia, o autor da aplicação precisa observar todo o conteúdo da mídia buscando todos os momentos que praia aparece para então usar tais momentos para especificar a sincronização dos efeitos sensoriais. Essa tarefa é muito custosa e pode induzir a erros [4].

Uma forma de diminuir o esforço de autoria de uma aplicação com efeitos sensoriais é utilizar uma abordagem declarativa. Assim, obtém-se uma clara separação entre a definição do efeito e detalhes de sua implementação. Existem propostas para a especificação de efeitos sensoriais em linguagens declarativas [3, 4], porém tal especificação necessita que o autor enderece diretamente os atuadores responsáveis por renderizar um efeito sensorial. Um ponto

negativo dessa proposta é que o código declarativo fica fortemente acoplado aos atuadores usados. Assim, diminui-se a portabilidade de tais aplicações, ou seja, que uma aplicação possa ser usada em outros ambientes sem que seja necessária uma alteração no seu código.

Por fim, uma forma de apoiar o processo de autoria de conteúdo *mulsemedia* é utilizar ferramentas de autoria. Tais ferramentas podem permitir uma rápida prototipação da aplicação sendo criada, como é o caso com simuladores [10]. No contexto de *mulsemedia*, entretanto, é importante que tal ferramenta dê suporte a execução de aplicações interativas e contendo efeitos sensoriais.

2 OBJETIVOS

Diante do grande interesse pelo aumento da imersão, em particular pela inclusão de efeitos sensoriais em aplicações multimídia e a dificuldade inerente do modelo de desenvolvimento sendo usado para criar tais aplicações, o principal objetivo da dissertação foi facilitar a autoria, sincronização e execução de efeitos sensoriais em aplicações multimídia interativas.

Uma abordagem para a autoria de efeitos sensoriais em aplicações interativas é reutilizar linguagens de autoria multimídia tradicionais, estendendo-as para este propósito. Dessa forma se ganha toda a capacidade de definição da sincronização de tais linguagens. Em particular, este trabalho foca na linguagem *Nested Context Language* (NCL) [7].

Em NCL, uma abordagem usual para definir a sincronização entre objetos de mídia é de duas formas. Primeiro são criadas âncoras temporais, indicando partes da mídia que são de interesse para a sincronização da aplicação. Posteriormente são criados relacionamentos que associam a ocorrência de eventos dessas âncoras a eventos em outros objetos de mídia na aplicação. Neste trabalho propomos o conceito de âncoras abstratas, cujo objetivo é aprimorar o processo de autoria de aplicações *mulsemedia* ao permitir a definição da sincronização de efeitos sensoriais com o conteúdo de mídias da aplicação. A partir de tal definição é feita a geração automática da sincronização por um software de reconhecimento que verifica dentro de cada mídia os momentos de apresentação dos componentes de cena relacionados a efeitos sensoriais.

Redes neurais podem ser utilizadas para realizar o reconhecimento de componentes cena relacionados a efeitos sensoriais. Porém, uma característica não explorada é o reconhecimento de efeitos sensoriais em ambas as modalidades de áudio e vídeo em conjunto. Neste trabalho, é proposta uma arquitetura de rede neural que realiza o reconhecimento de componentes de cena em ambas a modalidades conjuntamente.

NCL tem a vantagem de não levar em consideração o tipo de mídia que ela representa. As mídias em NCL podem ser do tipo áudio, vídeo, imagem, *scripts*, e mídias customizadas. De forma a permitir a especificação de efeitos sensoriais em aplicações NCL, este trabalho propõe ainda uma representação de tais efeitos similar à de objetos de mídia convencionais da aplicação. Assim, toda a capacidade de representação da sincronização e da interatividade de NCL pode ser reusada para a especificação de aplicações *mulsemedia*. A representação de um efeito sensorial utiliza uma extensão do método de posicionamento do padrão MPEG-V [6] para permitir a definição de coordenadas esféricas. Tais coordenadas permitem uma representação de posicionamento de efeitos sensoriais com maior precisão. Neste trabalho é apresentado um módulo que traduz tais descrições em NCL para descrições compatíveis no padrão MPEG-V e as envia pela rede.

Além de permitir a descrição de aplicações com efeitos sensoriais em linguagens multimídia interativas, este trabalho propõe um simulador 3D que permite visualização da ativação de efeitos sensoriais. O simulador recebe comandos de ativação pela rede e, desta forma, permite a visualização da execução de efeitos sensoriais integrada com a reprodução de aplicações interativas.

3 CONTRIBUIÇÕES E RESULTADOS

Esta seção apresenta as principais contribuições e resultados da dissertação. As contribuições são divididas em três frentes, que serão discutidas a seguir.

3.1 Autoria semiautomática da sincronização baseada no conteúdo de mídia

Uma âncora abstrata permite definir a sincronização de uma aplicação baseada em componentes de cena presentes no conteúdo de um objeto de mídia audiovisual. Ou seja, âncoras abstratas oferecem uma forma de descrever e sincronizar o conteúdo que não se encontra no documento, mas sim dentro de um objeto de mídia em si. Âncoras abstratas facilitam a autoria de aplicações com efeitos sensoriais por evitar que o autor precise identificar tempos de exibição de tais efeitos manualmente. Este processo é feito automaticamente de acordo com a identificação de componentes de cena em um objeto de mídia.

Foi feita a extensão da linguagem de autoria multimídia NCL de forma a permitir definição de âncoras abstratas. Esta extensão permite integrar âncoras abstratas com as definições de relacionamentos em NCL para reprodução de efeitos sensoriais em uma aplicação multimídia. NCL fornece o elemento `media` para definir nós que representam objetos de mídia. A Listagem 1 apresenta um exemplo de especificação de mídia com âncoras abstratas em NCL. Para permitir a definição de âncoras abstratas, a linguagem NCL foi estendida de forma que os elementos `area` (âncora em NCL) possam definir um novo atributo `tag`.

```
<media id="video1" src="video.mp4">
    <area tag="mar"/>
</media>
```

Listing 1: Exemplo de especificação de mídia com âncoras abstratas em NCL

Na Listagem 1, uma definição de âncora abstrata é feita com a tag `mar`. Isto é, esta âncora abstrata deve ser instanciada (e gerar âncoras comuns de NCL) em todos os momentos que `mar` é encontrado no conteúdo `video.mp4`. A Listagem 2 apresenta a definição em NCL após a etapa de instanciação. No exemplo da Listagem, a âncora abstrata `mar` foi instanciada em diversos momentos ao longo do vídeo. Importante notar que em NCL, os conectores definem uma relação geral que é instanciada por links para um determinado conjunto de participantes. Portanto sempre que uma âncora

[0]É importante notar que tal sincronização tem sempre uma única mídia como referência.

abstrata é instanciada, também são instanciados links que estejam relacionados a ela.

```
<media id="video1" src="video.mp4">
    <area id="mar_1" begin="01s" end="09s"/>
    <area id="mar_2" begin="17s" end="19s"/>
    <area id="mar_3" begin="23s" end="28s"/>
    <area id="mar_4" begin="60s" end="80s"/>
</media>
```

Listing 2: Resultado da etapa de instanciação de âncoras definidos no exemplo da Listagem 1

Para permitir a instanciação de âncoras abstratas em NCL, foi criado um processador de âncoras abstratas (do inglês *Abstract Anchor Processor* - AAP). A ferramenta faz interface com uma rede neural para reconhecer componentes de cena no conteúdo dos objetos de mídia de uma aplicação multimídia. De posse dos componentes de cena reconhecidos ao longo do tempo, o AAP realiza a instanciação das âncoras abstratas e dos relacionamentos associados a essas âncoras para definir a sincronização da aplicação como um todo. A ferramenta foi desenvolvida de forma a ser extensível e permitir a fácil integração com diversos softwares de reconhecimento. A Figura 2 apresenta uma visão em alto nível da rede neural construída. Em que um vídeo é separado em quadros que são levados a uma rede neural especializada em reconhecimento de imagens. O som do vídeo é transformado uma representação do espectro sonoro (espectrograma) e levados a uma segunda rede neural especializada em reconhecimento de áudio. Por fim a saída de ambas a redes neurais são comparadas por uma terceira rede neural que realiza a fusão dos dados e ao fim indica qual efeito sensorial está presenta naquele instante.

3.2 Aprimoramento do reconhecimento de componentes de cena em conteúdo audiovisual

De forma a suportar um reconhecimento mais preciso de componentes relacionados a efeitos sensoriais, foi proposta de uma arquitetura de rede neural bimodal. Tal proposta baseia-se na concepção de que as modalidades de áudio e vídeo podem ser usadas simultaneamente para identificar os momentos em que algum componente de cena é apresentado em uma mídia audiovisual. Portanto a arquitetura bimodal apresentada combina as modalidades para o reconhecimento de componentes de cena. Resultados experimentais indicam que a arquitetura de rede neural bimodal aprimora o reconhecimento de componentes de cenas relacionados a efeitos sensoriais.

Para treinar a rede neural foi criado um conjunto de dados bimodal. Para criação deste conjunto foi feito um método de adaptação do conjunto de dados AudioSet[1], que inicialmente contém apenas dados de áudio. Por fim resultando num conjunto de dados com exemplos de treinamento bimodais (i.e., de áudio e vídeo) que representam componentes de cena relacionados a efeitos sensoriais.

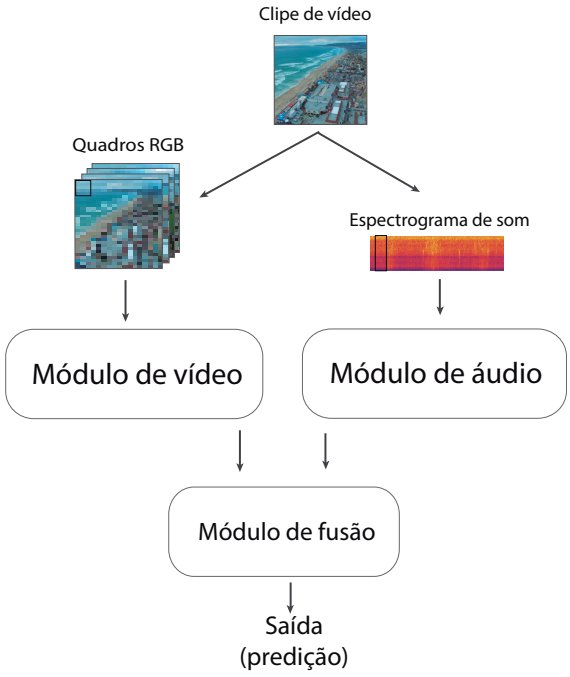

Figure 2: Ilustração da arquitetura de rede neural bimodal

3.3 Representação de efeitos sensoriais em aplicações interativas

Para suportar a criação de aplicações *mulsemedia* em NCL, foi desenvolvido um módulo de reprodução de efeitos sensoriais em NCL. O módulo de reprodução permite a criação de aplicações *mulsemedia* apoiando-se nas abstrações conceituais da linguagem NCL. O módulo, implementado em Lua [5], é capaz de interpretar eventos gerados pelo interpretador da linguagem NCL e traduzi-los em descrições de comandos de efeitos sensoriais compatíveis com o padrão MPEG-V. Os comandos gerados são enviados pela rede local para que reprodutores compatíveis com o padrão MPEG-V possam recebê-los e renderizar os efeitos descritos. O módulo facilita a autoria e reprodução de uma aplicação *mulsemedia* interativa ao integrar suas definições em uma linguagem multimídia orientada a eventos.

Afim de facilitar a especificação de posicionamento espacial de efeitos sensoriais, foi feita uma extensão do método de posicionamento de efeitos sensoriais do padrão MPEG-V. Esta extensão permite o uso de coordenadas esféricas, além das coordenadas tradicionais do MPEG-V, para definir a localização da ativação de efeitos sensoriais. Assim é permitindo ao autor um maior controle sobre o posicionamento de um efeito.

Por fim, foi desenvolvido um simulador 3D de efeitos sensoriais. O simulador permite a fácil prototipação de uma aplicação contendo efeitos sensoriais. O simulador foi construido em cima de software de modelagem 3D existente[1] e é capaz de representar uma sala tridimensional com atuadores de efeitos sensoriais e ativá-los

[1]Maxon Cinema 4D : https://www.maxon.net/cinema-4d

de acordo com descrições do padrão MPEG-V. O simulador também é capaz de apresentar o posicionamento de efeitos seguindo a extensão do método de posicionamento MPEG-V proposta neste trabalho. O simulador 3D facilita a autoria de aplicações *mulsemedia* interativas ao receber comandos e executá-los em tempo real. Além disso, o simulador 3D permite que um autor adicione e movimente atuadores existentes em um ambiente tridimensional.

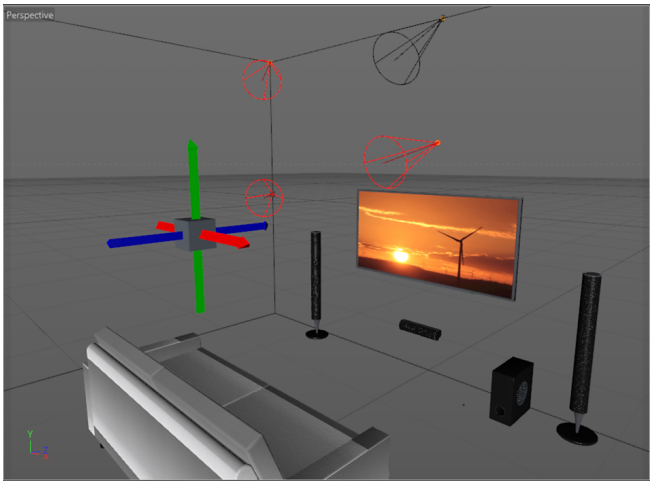

Figure 3: Ilustração da interface da simulação de efeitos sensoriais em 3D

4 PRODUTOS DA DISSERTAÇÃO

Um principal produto da dissertação para a linguagem NCL é *Abstract Anchor Processor* NCLAAP (https://github.com/GPMM/Abstract-Anchor). Também produto desta dissertação é o Tradutor de Efeitos sensoriais em NCL (https://github.com/GPMM/TES). Outro produto é simulador 3D (disponível sob pedido).

Para permitir o reconhecimento de efeitos sensoriais de conteúdo de mídia audiovisual, os produtos são a rede neural bimodal e conjunto de dados bimodal de efeitos sensoriais para treinamento de redes neurais (Ambos disponíveis em: https://github.com/MLRG-CEFET-RJ/bimodal_audioset).

Os seguintes artigos são produtos diretos desta dissertação:

- Raphael Abreu and Joel dos Santos.*Using Abstract Anchors to Aid the Development of Multimedia Applications with Sensory Effects*, no *ACM Symposium on Document Engineering* (*DocEng*) 2017.
- Raphael Abreu and Joel dos Santos. *Using Abstract Anchors for Automatic Authoring of Sensory Effects Based on Ambient Sound Recognition*, no Simpósio Brasilerio de Sistema Multimídia e Web (*WebMedia*) 2017.
- Raphael Abreu, Joel dos Santos, and Eduardo Bezerra.*A Bimodal Learning Approach to Assist Multi-sensory Effects Synchronization*, na International Joint Conference on Neural Networks (*IJCNN*) 2018.

Os seguinte trabalho é fruto da cooperação entre CEFET/RJ e grupo de pesquisa na Universidade Federal Fluminense durante o mestrado:

- Marina Josué, Raphael Abreu, Fábio Barreto, Douglas P. Mattos, Glauco F. Amorim, Joel A. F. dos Santos and Débora C. Muchaluat-Saade. *Modeling sensory effects as first-class entities in multimedia applications*, no *ACM Multimedia Systems Conference* (*MMSys*) 2018.

Destaca-se que WebMedia é a principal conferência de multimídia e hypermídia do Brasil. Enquanto as conferências DocEng e MMSys estão entre as principais internacionais. Destaca-se também que a conferência IJCNN está entre as principais conferências de redes neurais internacionais.

REFERENCES

[1] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*. New Orleans, LA.

[2] Gheorghita Ghinea, Christian Timmerer, Weisi Lin, and Stephen R. Gulliver. 2014. Mulsemedia : State of the Art, Perspectives, and Challenges. *ACM Transactions on Multimedia Computing, Communications, and Applications* 11, 1s (2014), 1–23. https://doi.org/10.1145/2617994

[3] Álan L.V. Guedes, Roberto G. de Albuquerque Azevedo, Sérgio Colcher, and Simone D.J. Barbosa. 2016. Extending NCL to Support Multiuser and Multimodal Interactions. In *Proceedings of the 22Nd Brazilian Symposium on Multimedia and the Web (Webmedia '16)*. ACM, New York, NY, USA, 39–46. https://doi.org/10.1145/2976796.2976869

[4] Álan Lívio Vasconcelos Guedes, Roberto Gerson de Albuquerque Azevedo, and Simone Diniz Junqueira Barbosa. 2017. Extending multimedia languages to support multimodal user interactions. *Multimedia Tools and Applications* 76, 4 (2017), 5691–5720.

[5] Roberto Ierusalimschy. 2006. *Programming in lua.* Roberto Ierusalimschy.

[6] ISO/IEC 23005-3 2016. *Information technology – Media context and control – Part 3: Sensory information.* Standard. International Organization for Standardization, Geneva, CH.

[7] ITU. 2009. Nested Context Language (NCL) and Ginga-NCL for IPTV services. http://www.itu.int/rec/T-REC-H.761-200904-S. ITU-T Recommendation H.761.

[8] Marina Josué, Raphael Abreu, Fábio Barreto, Douglas P. Mattos, Glauco F. Amorim, Joel A. F. dos Santos, and Débora C. Muchaluat-Saade. 2018. Modeling Sensory Effects as First-Class Entities in Multimedia Applications. In *ACM Multimedia Systems Conference*.

[9] Christian Timmerer, Markus Waltl, Benjamin Rainer, and Hermann Hellwagner. 2012. Assessing the quality of sensory experience for multimedia presentations. *Signal Processing: Image Communication* 27, 8 (2012), 909–916. https://doi.org/10.1016/j.image.2012.01.016

[10] Markus Waltl, Benjamin Rainer, Christian Timmerer, and Hermann Hellwagner. 2013. An end-to-end tool chain for Sensory Experience based on MPEG-V. *Signal Processing: Image Communication* 28, 2 (2013), 136–150. https://doi.org/10.1016/j.image.2012.10.009