XXIX Congresso de Iniciação Científica UNICAMP 2021

Visualização de Componentes Principais em Realidade Virtual

Huang Tzu Jan, Rodolfo Luis Tonoli, Paula Dornhofer Paro Costa
Depto. Eng. de Computação e Automação (DCA), Faculdade de Eng. Elétrica e de Computação (FEEC)
Universidade Estadual de Campinas (Unicamp)
Campinas, Brasil
e-mail: h198887@dac.unicamp.br, paulad@unicamp.br

***Resumo*—Nos últimos anos, o crescimento sem precedentes da produção de dados, com graus crescentes de complexidade, vem tornando a análise de dados uma tarefa cada vez mais desafiadora. Ao mesmo tempo, a Realidade Virtual vem se tornando cada vez mais popular, permitindo aos pesquisadores avaliarem como a imersão em ambientes de realidade virtual poderia agregar valor ao processo de análise visual de dados volumosos, também chamado de Visual Analytics (VA), explorando formas inovadoras de extração de conhecimento dos dados. O framework ImmVIS, desenvolvido na Faculdade de Engenharia Elétrica e de Computação da Unicamp, foi projetado para permitir a exploração de diferentes paradigmas de visualização de dados em realidade virtual, possibilitando a integração de diferentes plataformas de visualização a serviços de análise de dados em Python. O presente trabalho propôs a integração de uma nova funcionalidade de visualização baseada na Análise de Componentes Principais ao ImmVIS. A Análise de Componentes Principais é uma análise estatística que permite encontrar os eixos de maior covariância de um conjunto de dados multidimensional, podendo ser utilizada para guiar a visualização de dados volumosos.**

***Palavras-chave*—Visual Analytics, ciência dos dados, visualização em realidade virtual**

I. Introdução

Nos últimos anos, os dados vêm sendo gerados em volumes e velocidades sem precedentes, levando a novos desafios no processo de análise dos mesmos. A área de *Visual Analytics* dedica-se a desenvolver técnicas de manuseio de volumes massivos, heterogêneos e dinâmicos de informações, integrando o julgamento humano a algoritmos automáticos que propõem representações visuais dos dados e possibilitam uma análise interativa. A área de *Visual Analytics* agrega diferentes áreas de pesquisa, incluindo visualização, mineração de dados e estatística [1]–[3].

Como uma ferramenta de auxílio para VA, a Realidade Virtual (RV) é uma tecnologia emergente que apresenta potencial em processamento de dados multidimensionais e aquisição do conhecimento, uma vez que há a possibilidade de visualização e exploração colaborativa dos dados fora dos ambientes convencionais, que se limitam ao 2D, aumentando a interação com usuário ao deixá-lo imerso em um ambiente virtual 3D, com efeitos estereoscópicos e rastreamento de movimentos do corpo do usuário, proporcionando um entendimento mais intuitivo e a uma melhor retenção das relações e *insights* extraídos dos dados [4], [5].

Uma das formas de criar aplicações em RV é utilizando motores gráficos de jogos (*game engines*) como o Unity, uma plataforma projetada para o desenvolvimento de aplicações de entretenimento, mas que também tem sido muito utilizada para desenvolvimentos em Realidade Virtual. Em particular, o Unity foi a plataforma escolhida para a implementação do ImmVIS, um framework *open source* desenvolvido com o objetivo de possibilitar a implementação de variadas funcionalidades de Immersive Analytics, e que adota um protocolo de comunicação entre serviços (gRPC) que permite conectar rotinas de análise de dados de diferentes plataformas e linguagens de programação, tais como scripts de análise de dados escritos em Python, ao ambiente de RV [6].

Assim, o presente trabalho teve como objetivo adicionar uma nova funcionalidade de análise ao ImmVIS, baseada na Análise de Componentes Principais, que permite encontrar as dimensões de maior variância de um conjunto de dados multidimensional.

II. Método

A. PCA

A Análise dos Componentes Principais, ou simplesmente PCA, do inglês *Principal Component Analysis*, é uma técnica amplamente utilizada para a redução de dimensionalidade, compressão dos dados, extração de atributos e visualização de dados. A análise é feita encontrando a projeção linear dos dados em um espaço de dimensão reduzida que faça com que as variáveis obtidas, denominadas de componentes principais, possuam máxima variância explicada (informação), podendo ser extraídas do PCA informações como a variância explicada

Este trabalho foi financiado pelo Programa Institucional de Bolsas de Iniciação Científica (PIBIC), CNPq.

de cada componente, a razão dessa variância, a covariância dos dados, utilizando a biblioteca *sklearn*, uma biblioteca *open source* da linguagem Python, utilizada para análise preditiva dos dados.

O módulo desenvolvido no ImmVIS permite, após a seleção do conjunto de dados a ser explorado e o tipo de análise (PCA), a escolha da quantidade de componentes principais a serem analisadas, podendo ser estudados gráficos 1D, 2D e 3D.

B. Dataset Iris

Para efeito de ilustração, foi escolhido o Dataset Iris, um conjunto de dados que descreve três espécies de flor Iris através de quatro atributos e cinco colunas: comprimento da sépala, largura da sépala, comprimento da pétala, largura da pétala e espécie da flor. O dataset contempla três tipos de espécies da flor: Iris setosa, Iris versicolor e Iris virginica.

A aplicação do PCA sobre o dataset pode reduzir a dimensionalidade (4D para 3D ou 2D, por exemplo) de forma a permitir a representação desse conjunto de dados no mundo virtual, proporcionando novas interpretações dos dados em diferentes dimensões analisadas.

O PCA pode ser usado para lidar com conjunto de dados de dimensões maiores como imagens, que podem ter milhares de colunas, e aplicando PCA a quantidade de colunas pode ser reduzida a centenas ou dezenas sem grandes perdas de informações, pois PCA maximiza a variância para a primeira componente principal e para as próximas componentes com as variâncias restantes, retendo a máxima quantidade de informações ao reduzir a dimensionalidade do conjunto de dados.

III. Resultados

O resultado da implementação pode ser visualizado nas Figuras 1 e 2, onde os eixos em vermelho, verde e azul são a primeira, segunda e terceira componentes principais, respectivamente. As figuras mostram três grupos de pontos, um para cada espécie de Iris, da esquerda para direita: Iris setosa, Iris versicolor e Iris virginica.

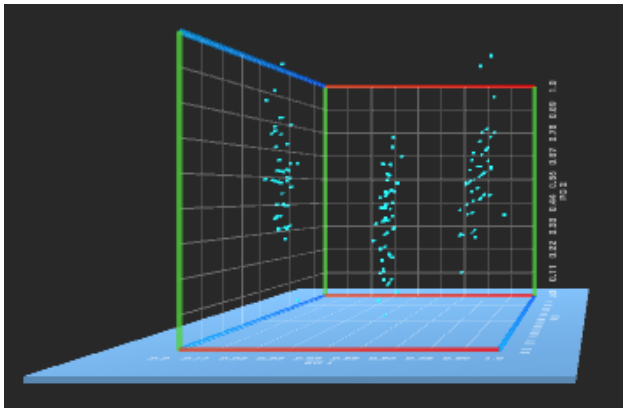

Figura. 1. Scatterplot do Dataset Iris usando 3 componentes principais, onde da esquerda para direita: Iris setosa, Iris versicolor e Iris virginica.

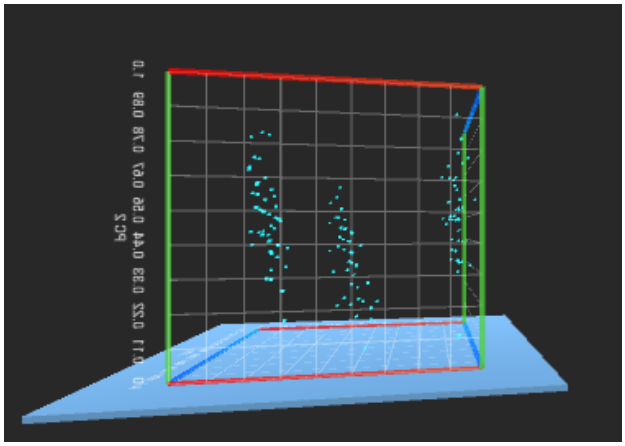

Figura. 2. Scatterplot do Dataset Iris usando 3 componentes principais, com rotação de 168º, onde da direita para esquerda: Iris setosa, Iris versicolor e Iris virginica.

IV. Conclusão

O trabalho agregou uma funcionalidade de PCA ao ImmVIS que apresenta potencial de ajudar na visualização dos dados para entender as relações entre eles, uma vez que os dados estão sendo gerados cada vez mais complexos e multidimensionais, o que torna a análise de dados uma tarefa difícil em gráficos 2D.

Neste trabalho, ainda há pontos a serem melhorados, como apresentar ao usuário as informações que PCA fornece, como a matriz de covariância e variância explicada cumulativa (a variância explicada informa quanta informação pode ser atribuída a cada uma das componentes principais). Por fim, expandir novas funcionalidades para o framework, e, porventura, realizar avaliações perceptuais das visualizações.

Referências

[1] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler, "Visual analytics: Scope and challenges," in *Visual data mining*. Springer, 2008, pp. 76–90.

[2] P. C. Wong and J. Thomas, "Visual analytics," *IEEE Computer Graphics and Applications*, no. 5, pp. 20–21, 2004.

[3] D. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, "Mastering the information age: solving problems with visual analytics," 2010.

[4] C. Donalek, S. G. Djorgovski, A. Cioc, A. Wang, J. Zhang, E. Lawler, S. Yeh, A. Mahabal, M. Graham, A. Drake *et al.*, "Immersive and collaborative data visualization using virtual reality platforms," in *2014 IEEE International Conference on Big Data (Big Data)*. IEEE, 2014, pp. 609–614.

[5] B. Laha and D. A. Bowman, "Identifying the benefits of immersion in virtual reality for volume data visualization," in *Immersive visualization revisited workshop of the IEEE VR conference*. Citeseer, 2012, pp. 1–2.

[6] F. Pedroso and P. D. P. Costa, "Improved knowledge from data: Building an immersive data analysis platform," in *20th Symposium on Virtual and Augmented Reality, 2018*, 2018.