Simpósio de Integração Acadêmica
"Bicentenário da Independência: 200 anos de ciência, tecnologia e inovação no Brasil e 96 anos de contribuição da UFV"
SIA UFV 2022
UFV Universidade Federal de Viçosa

ANÁLISE DA VULNERABILIDADE HUMANA ACERCA DA IDENTIFICAÇÃO DE DEEPFAKES

Universidade Federal de Viçosa – Departamento de Engenharia Elétrica

Denise Cristina Henrique de Freitas (denise.henrique@ufv.br), Kétia Soares Moreira (ketia@ufv.br) e Helena Cristo Martins (helena.martins@ufv.br)

Palavras-Chave: Deepfake, Inteligência Artificial, base de dados

Área temática: Ciências Exatas e Tecnológicas | Grande área: Engenharia Elétrica

Projeto de pesquisa

Introdução

Deepfake consiste em uma tecnologia que utiliza Inteligência Artificial (IA), mais especificamente a aprendizagem profunda, para criar vídeos ou imagens falsas, com adulterações que podem passar despercebidas aos olhos humanos. Essa técnica, que já gerou desde conteúdos pornográficos com celebridades até discursos fictícios de políticos, cresceu tanto nos últimos anos, que hoje se faz necessário estudos acerca das metodologias usadas para sua criação, de modo a entender o seu funcionamento, e assim, possibilitar o desenvolvimento de ferramentas para combater essa disseminação de informações falsas.

Objetivo

Criar uma base de dados de *deepfakes* por meio de estudos metodológicos e, posteriormente, avaliar a vulnerabilidade humana na identificação das mesmas.

Material e Métodos

Neste trabalho utilizou-se o método de criação de *deepfakes* proposto no artigo "First Order Motion Model for Image Animation"[2]. Assim, através da implementação de um código em Python, foi criada uma base de dados de vídeos adulterados. Conforme apresentado na Figura 1, esses vídeos são compostos, da esquerda pra direita, pela foto fonte, pelo vídeo original e pelo vídeo manipulado.

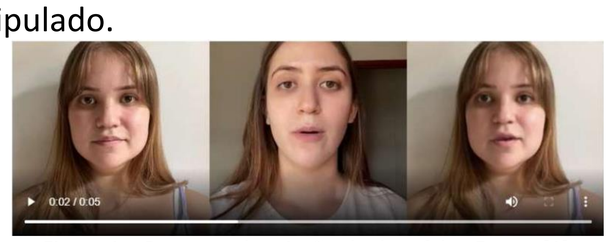

Figura 1 - Processo de criação da deepfake. Fonte: [1].

Posteriormente, realizou-se análises, tanto objetiva, com a utilização da métrica PSNR (Peak Signal to Noise Ratio); como subjetiva, utilizando a métrica MOS (Mean Opinion Score). A PSNR relaciona a entrada e saída de compressão de perdas que avalia se houve ou não ruído introduzido na imagem ou frames (um vídeo com qualidade satisfatória apresenta um valor médio do PSNR entre 25 e 31 dB). Já para obtenção da métrica MOS, foi enviado um formulário a dois grupos de pessoas, no qual elas deveriam atribuir notas de 1 a 5 para a qualidade e veracidade dos vídeos, um desses grupos conheciam as pessoas dos vídeos e o outro não.

Apoio Financeiro

CNPq

Resultados e Discussão

Tabela 1 – Avaliação dos vídeos pela métrica MOS

	Grupo 1		Grupo 2	
Identificação do Vídeo	Média em relação a veracidade	Média em relação a qualidade	Média em relação a veracidade	Média em relação a qualidade
1	4,64	4	4	4,4
2	3,91	4,18	4	4,6
3	3,82	4,18	4,4	4,6
4	3,27	3,73	3,6	4,8
5	3,64	4	3,8	4,6
6	3,82	3,82	3,6	4,6
7	3,64	3,73	4,6	4,8
8	2,64	3,45	3	4,4
9	4,36	4,1	4	4,8
10	4,36	4,27	4,4	4,8
Média geral	**3,71**	**3,95**	**3,94**	**4,64**

Conforme apresentado na tabela 1, os resultados obtidos na análise subjetiva foram satisfatórios, uma vez que, um valor em torno de 4,3 e 4,5 é considerado um avaliação de excelente qualidade, enquanto algo inaceitável estaria abaixo de um MOS de aproximadamente 3,5. Logo, no grupo 1 obteve-se ambos os aspectos (veracidade e qualidade) dentro de algo considerado aceitável, enquanto que para o grupo 2, a veracidade foi classificada como razoável/aceitável enquanto que a qualidade foi considerada excelente.

Em relação a análise objetiva foram avaliados 4 frames de cada um dos vídeos criados e, todos os resultados obtidos foram acima de 37 dB, se enquadrando como excelentes.

Conclusões

As análises realizadas nesse projeto mostraram que, mesmo sem a utilização de recursos computacionais avançados, é possível criar *deepfakes* suficientemente convincentes para enganar muitas pessoas. Assim, esse trabalho cumpriu com êxito os objetivos propostos, evidenciando a vulnerabilidade humana acerca da percepção e julgamento dos *deepfakes* criados. Posto isso, torna-se evidente a importância do desenvolvimento de ferramentas eficazes para identificação e combate dessas falsificações.

Bibliografia

[1] MARTINS, Helena Cristo. **Compreensão da vulnerabilidade humana na identificação de *deepfakes***. Orientador: Kétia Soares Moreira. 2022. 48 p. Trabalho de conclusão de curso (Bacharelado em Engenharia Elétrica) - Universidade Federal de Viçosa, Viçosa, 2022.

[2]SIAROHIN, Aliaksandr et al. First order motion model for image animation. Advances in Neural Information Processing Systems, v. 32, 2019.