MELHORIA DE QUALIDADE DE VÍDEO COMPRIMIDO ATRAVÉS DO USO DE REDES NEURAIS PROFUNDAS

GILBERTO KREISLER[1]; GARIBALDI DA SILVEIRA JUNIOR[2]; BRUNO ZATT[3]; DANIEL PALOMINO[4]; GUILHERME CORRÊA[5]

[1]*Universidade Federal de Pelotas – gkfneto@inf.ufpel.edu.br*
[2]*Universidade Federal de Pelotas – garibaldi.dsj@inf.ufpel.edu.br*
[3]*Universidade Federal de Pelotas – dpalomino@inf.ufpel.edu.br*
[4]*Universidade Federal de Pelotas – zatt@inf.ufpel.edu.br*
[5]*Universidade Federal de Pelotas – gcorrea@inf.ufpel.edu.br*

1. INTRODUÇÃO

O volume de dados provenientes de vídeos digitais tem crescido cada vez mais na internet. Com a pandemia de COVID-19, empresas com atividades relacionadas a vídeos digitais como TikTok, NetFlix e YouTube tiveram um grande aumento de demanda. Só durante o primeiro mês da pandemia, o volume de dados proveniente da transmissão de vídeos pela internet aumentou em 32,6% (STATISTA, 2022). Além disso, vídeos de alta resolução, como *Ultra-High Definition* (UHD) 4K e 8K, têm se tornado cada vez mais comuns. De acordo com (CISCO, 2020), até o fim do ano de 2023, vídeos em 4K representarão 66% do consumo de internet por aparelhos de televisão, sendo este percentual maior do que o previsto em 2018 (33%).

Para transmitir estes vídeos sem nenhum tipo de compressão é necessária uma alta largura de banda, porém este é um recurso limitado. Dessa forma a compressão de vídeos é um processo necessário para que um vídeo seja transmitido sem consumir muita largura de banda e com uma qualidade aceitável. Porém, o processo de compressão acaba por introduzir artefatos nos vídeos que degradam a Qualidade de Experiência (*Quality of Experience* – QoE). Dessa forma é necessária a adição de filtros para processar o vídeo após a descompressão.

Grande parte das arquiteturas propostas para o problema de melhoria de qualidade de vídeo (*Video Quality Enhancement* - VQE) utiliza apenas vídeos comprimidos conforme o padrão HEVC, tanto para treinamento como teste, não analisando a efetividade dos modelos gerados para a melhoria de vídeos comprimidos por outros padrões de codificação. Este é o caso da arquitetura *Spatio-Temporal Deformable Fusion* (STDF), proposta por DENG ET. AL.(2020), que adota um alinhamento de características utilizando convoluções deformáveis aplicadas a múltiplos quadros ao invés do típico processo de estimação e compensação de movimento adotado por outras arquiteturas, como a *Multi Frame Quality Enhancement* (MFQE) de YANG ET. AL.(2018).

Neste trabalho, apresentamos um novo modelo baseado na arquitetura STDF (DENG ET. AL., 2020). Esta proposta, denominada multi-codec, é realizada usando um *dataset* misto composto por vídeos comprimidos pelos codecs VVC e AV1. Resultados experimentais mostram que o modelo proposto alcança um aumento de qualidade objetiva consistente para vídeos comprimidos com múltiplos codecs, atingindo um valor de ΔPSNR até 0,382 dB.

2. METODOLOGIA

O processo de treinamento do modelo STDF multi-codec começa com a divisão do *dataset* de vídeo em duas partes iguais: 54 vídeos comprimidos usando VVC com QP 37 e 54 vídeos comprimidos usando AV1 com CQ 55. O *VVC Test Model* (VTM) (versão 13.0) foi utilizado como software de referência para todas as codificações de VVC, seguindo a configuração temporal *Low Delay*. O software de referência libaom (*hashcode* 3.3) foi utilizado para todas as codificações AV1. A divisão dos vídeos foi feita de acordo com o tipo de resolução.

A partir dessa divisão, foi gerado o *dataset* multi-codec, utilizado para treinamento o modelo STDF (DENG ET. AL., 2020). Durante o processo de treinamento, o modelo é alimentado com pares de imagens, sendo uma imagem de referência não compactada e outra imagem compactada pelo VVC ou AV1. O objetivo do treinamento é fazer com que o modelo aprenda a mapear as imagens compactadas de volta para as imagens originais de alta qualidade. O processo de treinamento é repetido várias vezes, ajustando os pesos do modelo para reduzir a diferença entre as imagens geradas pelo modelo e as imagens originais de alta qualidade. O modelo treinado é então avaliado usando um conjunto de dados de validação e testado em vídeos comprimidos com outros codecs e configurações de quantização.

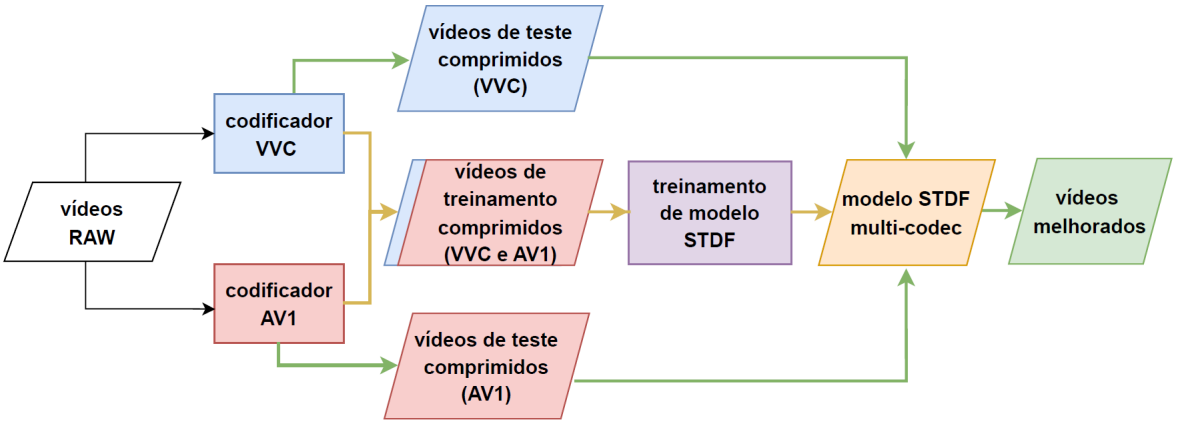

Figura 1. Metodologia de treinamento e teste do modelo STDF multi-codec.

O processo de treinamento é representado pelo caminho com setas amarelas na Figura 1, enquanto a etapa de testes é representada pelo caminho com setas verdes. O treinamento foi realizado usando a implementação de referência STDF (XING., DENG., 2020). Como os dois módulos do STDF são baseados em CNN, a arquitetura é unificada e pode ser treinada de ponta a ponta. Para o treinamento, foi utilizado um computador com a seguinte configuração: processador *AMD Ryzen 7 5700X*; 32 GB de memória RAM; GPU *Nvidia Geforce RTX 3070* com 8 GB de VRAM. Os parâmetros referentes ao tamanho do lote (*batch size*) e número de iterações foram adaptados para atingir o mesmo número de épocas usado em DENG ET. AL.(2020) com uma única GPU (ou seja, um tamanho do lote de 8 e 300.000 iterações).

3. RESULTADOS E DISCUSSÃO

A Tabela 1 apresenta os resultados de VQE obtidos para o modelo STDF multi-codec proposto considerando as 18 sequências de vídeo de teste. Os resultados são apresentados para as 8 versões dos vídeos de teste, ou seja,

considerando os quatro codecs e duas configurações de quantização. Os vídeos de teste são agrupados por dimensão de acordo com a sua classe (BOYCE ET. AL. 2018): Classe A: 2560x1600, Classe B: 1920x1080, Classe C: 832x480, Classe D: 416x240, Classe E: 1280x720. Os resultados são apresentados em PSNR, pois esta é a métrica mais comummente utilizada em trabalhos relacionados, permitindo comparações.

Tabela 1. Resultados do modelo STDF multi-codec para vídeos comprimidos com diferentes codecs.

Dataset de treino		ΔPSNR(dB)							
		HEVC		VVC		VP9		AV1	
		QP 32	QP 37	QP 32	QP 37	CQ 43	CQ 55	CQ 43	CQ 55
Classe A	*Traffic*	0,314	0,291	0,193	0,190	0,277	0,358	-0,012	0,166
	PeopleOnStreet	0,37	0,368	0,130	0,149	0,340	0,391	0,055	0,19
Classe B	*Kimono*	0,249	0,193	0,114	0,112	0,217	0,195	0,093	0,12
	ParkScene	0,15	0,105	0,115	0,078	0,183	0,164	0,123	0,155
	Cactus	0,244	0,239	0,123	0,163	0,199	0,230	0,011	0,095
	BQTerrace	0,14	0,198	-0,009	0,052	0,074	0,192	-0,127	0,032
	BasketballDrive	0,313	0,316	0,088	0,152	0,261	0,317	-0,001	0,16
Classe C	*RaceHorses*	0,254	0,246	0,084	0,113	0,244	0,283	0,66	0,176
	BQMall	0,388	0,342	0,211	0,248	0,314	0,375	0,013	0,221
	PartyScene	0,492	0,372	0,265	0,258	0,436	0,428	0,203	0,279
	BasketballDrill	0,447	0,443	0,008	0,149	0,424	0,442	0,031	0,234
Classe D	*RaceHorses*	0,348	0,298	0,177	0,175	0,352	0,328	0,22	0,261
	BQSquare	0,803	0,643	0,431	0,375	0,752	0,775	0,571	0,689
	BlowingBubbles	0,46	0,352	0,328	0,316	0,403	0,377	0,246	0,311
	BaskeetballPass	0,573	0,463	0,380	0,392	0,549	0,515	0,257	0,43
Classe E	*FourPeople*	0,514	0,431	0,316	0,342	0,450	0,514	-0,064	0,209
	Johnny	0,398	0,349	0,221	0,256	0,352	0,410	0,032	0,21
	KristenAndSara	0,428	0,384	0,226	0,260	0,345	0,461	-0,075	0,153
Média		**0,382**	**0,335**	**0,189**	**0,210**	**0,343**	**0,375**	**0,091**	**0,229**

Pode ser observado que o modelo multi-codec proposto obtém resultados médios positivos em todos os cenários. Em média, o modelo proposto é capaz de aumentar a qualidade objetiva da imagem entre 0,091 dB (AV1, CQ 43) e 0,382 dB (HEVC, QP 32). Além disso, também pode-se observar que o modelo foi capaz de melhorar a qualidade de vídeos comprimidos com outras configurações de quantização além daquelas usadas para comprimir os vídeos de treinamento, como é o caso do HEVC QP 32 (0,382 dB), VVC QP 32 (0,189 dB), VP9 CQ 43 (0,343) e AV1 CQ 43 (0,091).

A Tabela 2 apresenta resultados médios obtidos com os modelos STDF single-codec e, na última linha, os resultados obtidos com o modelo STDF multi-codec replicados. Em dois terços dos modelos single-codec, podem ser percebidos resultados negativos no teste, indicando que os modelos não são eficazes para todos os quatro padrões/formatos de codificação testados. De fato, em 6 dos 24 testes (25%) realizados com modelos single-codec, valores negativos (isto é, redução na qualidade visual) foram percebidos. O único modelo single-codec que gera resultados positivos em todos os casos de teste é aquele treinado com vídeos gerados pelo codec AV1. Ainda assim, o modelo multi-codec alcança melhores resultados do que este em praticamente todos os casos de teste, exceto para vídeos comprimidos usando o AV1, o que é esperado.

Tabela 2. Comparação entre o STDF multi-codec e o STDF single-codec

Dataset de treino	ΔPSNR (dB)							
	HEVC		VVC		VP9		AV1	
	QP 32	QP 37	QP 32	QP 37	CQ 43	CQ 55	CQ 43	CQ 55
HEVC QP 37	0,362	0,755	-0,217	0,250	-0,465	0,357	-1,479	-0,506
VVC QP 37	0,446	0,529	0,216	0,371	0,050	0,385	-0,530	-0,016
AV1 CQ 55	0,346	0,285	0,137	0,144	0,368	0,389	0,109	0,286
Multi-codec	**0,382**	**0,335**	**0,189**	**0,210**	**0,343**	**0,375**	**0,091**	**0,229**

4. CONCLUSÕES

Este artigo apresentou um novo modelo para a arquitetura STDF de VQE capaz de melhorar a qualidade visual de vídeos comprimidos com diferentes codecs de vídeo. O modelo apresentado propõe uma abordagem inovadora, que leva em consideração que diferentes padrões e formatos de codificação de vídeo introduzem diferentes tipos e níveis de artefatos de compressão na imagem. Dessa forma, o modelo foi treinado com vídeos gerados por múltiplos padrões de codificação de vídeo, para que pudesse aprender a reduzir esses artefatos de forma mais eficaz. Ao contrário dos trabalhos estado da arte, que geralmente são treinados com apenas um codec de vídeo, o modelo proposto neste artigo foi treinado com vídeos comprimidos pelos codecs VVC e AV1. Os resultados obtidos foram promissores, mostrando que o modelo foi capaz de melhorar a qualidade visual dos vídeos comprimidos em todos os casos testados.

5. REFERÊNCIAS BIBLIOGRÁFICAS

STATISTA. **Semiconductor market size worldwide from 1987 to 2020**. 2022. Acessado em 23 mar. 2022. Online. Disponível em: https://www.statista.com/statistics/266973/global-semiconductor-sales-since-1988/

CISCO. **Cisco annual internet report (2018–2023) white paper**. 2020. Acessado em 15 fev. 2023. Online. Disponível em: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

Xing, Q. and Deng, J. **PyTorch implementation of STDF**.2020. Acessado em 21 set. 2023. Online. Disponível em: https://github.com/ryanxingql/stdf-pytorch, version 1.0.0, 2020.

Boyce, J., Suehring, K., and Li, X. (2018). Jvet-j1010: **Jvet common test conditions and software reference configurations**. JVET-J1010

Deng, J., Wang, L., Pu, S., and Zhuo, C. (2020). **Spatio-temporal deformable convolution for compressed video quality enhancement**. Proceedings of the AAAI conference on artificial intelligence, v. 34, p. 10696–10703.

Yang, R., Xu, M., Wang, Z., and Li, T. (2018). **Multi-frame quality enhancement for compressed video**. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, p. 6664–6673.