INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC 1/SC 29/WG 4
MPEG VIDEO CODING

ISO/IEC JTC 1/SC 29/WG 4 m70724
October 2025, Geneve

Title: **[VCM] Improved neural-network-based chroma reconstruction**

Source: **Poznan University of Technology**
Olgierd Stankiewicz, Sławomir Różek, Mateusz Lorkiewicz, Hubert Żabiński, Tomasz Grajek, Sławomir Maćkowiak, Maciej Wawrzyniak, Jakub Stankowski, Marek Domański

Abstract

This documents presents an improved version of VCM tool proposal m70249 [1] which consists in neural-network-based chrominance reconstruction in the decoder, based solely on luminance, instead of chroma transmission. The difference between the previous version is change of color-format used for processing: Now VCM-native YCbCr is used, instead of CIE Lab previously used in [1]. Just like before, two alternative neural network models can be used in the decoder: a neural network based on ECCV16 [2] architecture and SIGGRAPH17 [3] architecture. The selection on network used is done adaptively the encoder and then transmitted in the bitstream. The usage of the propose tool provides gains of up to 22.57% of average bitrate reduction.

1. The idea of the tool

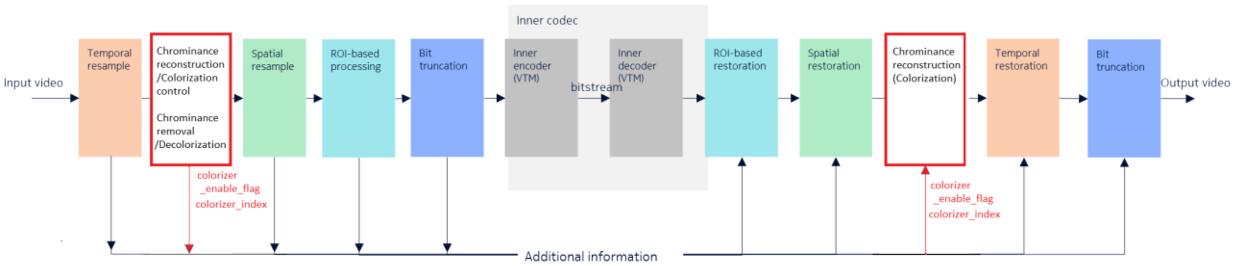

Fig 1. The general (idealistic) idea of the tool and its placement (red boxes) in the VCM pipeline- the same as in m70249 [1]

The main idea of the tool (Fig 1) remains the same as in [1] and consists in not sending chrominance directly, and reconstructing it at the decoder side (from luminance only) instead. This is done with the use of neural-network-based approach.
Two alternative neural network models are employed for this task: based on ECCV16 architecture [2] and SIGGRAPH17 architecture [3], implemented in open-source packages [4].
The decision about which network to use (and whether to enable chroma reconstruction tool or disable it) is done at the encoder. The selected network index (currently 0 or 1) is transmitted in the bitstream.

Just as mentioned in [1], this general (idealistic) in implementation is adapted to address two issues in VCM and VCM-RS:

- Bit truncation tool
  Bit Truncation tool offers possibility to conditionally shift luma component by given amount of bits to the right at the encoder, which corresponds to dimming the image.

Syntax of Bit Truncation tool allows signalization of reverse shift value, so that the luma component is shifted left in the decoder, which corresponds to increasing lightness.
Unfortunately, operation of BitTruncation tool is NOT symmetrical:

- BitTruncation tool is placed in different places in the processing pipeline (different in the encoder and in the decoder)
- The information which is indeed signalled does NOT correspond to restoring the original luminance dynamic range.
  E.g. the encoder performs shift of luminance by 1 bit to the right, but is NOT signalling 1 bit shift left information to the decoder. As a result, the video outputted from the VCMRS decoder is 2 times darker than the original. Consequently, the decoded dynamic range of luminance is twice smaller.

Due to the abovementioned BitTruncation tool mechanism, and resulting changed dynamic range of the luminance component in the decoder, the operation of the colorization could be degraded.
Therefore it is also proposed to signal "real" luma shift (luma_pre_shift) value as an additional field in the proposed bitstream syntax. This is used inside of the proposed tool to temporarily restore the original luminance dynamic range, just for the sake of chrominance restoration. This does not affect output from the decoder luminance, which is not affected. In particular, the proposed tool does not affect luminance at all.

- Descriptor calculation
  Spatial resample and ROI tool employ pre-calculated descriptors of objects. Those descriptors are calculated inside of the pipeline of VCMRS in the place of respective tools, e.g. inside Spatial Resample and ROI.
  Therefore it is not possible to perform decolorization (chrominance removal) prior to those tools, because it would affect calculation of the descriptors.
  In order to solve this issue, the decolorization step was moved to a separate box in the encoder .

The implemented general idea of the proposed tool, with fixes addressing the abovementioned issues, is presented in Fig. 2.

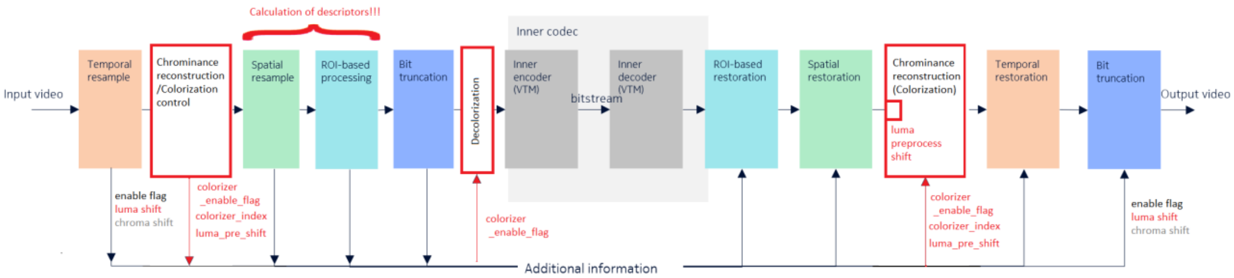

Fig 2. The implemented general idea of the tool and its placement (red boxes) in the VCM pipeline.

2. Operation at the encoder

The operation of the proposed tool in VCM encoder is presented in Fig. 3. The encoder makes decision about the selection of the neural network which will be used at the decoder for chrominance restoration. Consequently it decides whether to send unaltered video, of decolorized version (chrominance cleared, luminance unaltered).
The steps are as follows:

- First, decolorized (desaturated) version of the input video is calculated. This done by discarding chrominance information and preserving (unaltered) luminance.
- This is then used to check quality of chroma restoration with the use of two networks, based on ECCV architecture [2] and SIGGRAPH architecture [3].
- The quality is evaluated through comparison of the synthetic (generated) chrominance with the original chrominance information (in the input video), based on correlation Pearson coefficient and PSNR. The Pearson correlation is zeroed if PSNR is too low.
- Next, the better performing network is selected, but not worse than having Pearson correlation coefficient of 0.7.
- If the networks are underperforming (r<=0.7), the tool is disabled (colorizer_enable_flag=0) and the original video is fed for further encoding processes.

- If one of the networks is selected (r>07), the proposed colorization tool is enabled (colorizer_enable_flag =1) and the index of the selected network (colorizer_index) is transmitted to the decoder. In such a scenario, the decolorized video is fed for the further encoding prosses.
- Additionally, luma_pre_shift value is also signalled (the same value as really used right shift in BitTruncation tool).

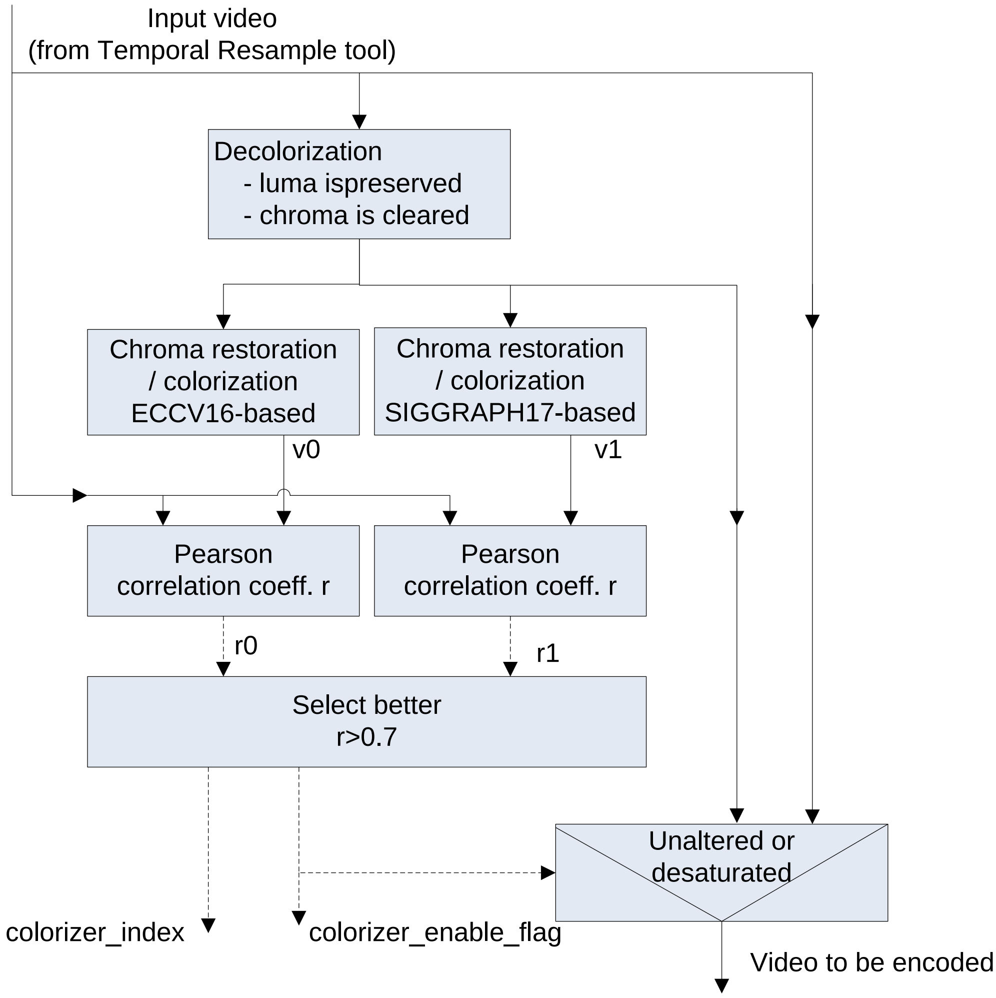

Fig. 3. Operation of the proposed tool in the encoder.

3. Syntax

8.3.5.5. Colorization syntax

colorization( ) {	Descriptor
colorizer_enable_flag	u(1)
if(colorizer_enable_flag) {	
colorizer_index	u(1)
colorizer_luma_pre_shift	u(3)
}	
byte_alignment( )	
}	

8.4.5.5. Colorization semantics

colorizer_enable_flag equal to 1 specifies that the colorization process is enabled. colorizer_enable_flag equal to 0 specifies that the colorization process is disabled.

colorizer_index – 1 bit – equal to 0 or 1, specifies the index of chroma restoration network to be in colorization process.

colorizer_luma_pre_shift equal to 0, 1, 2, 3, 4, 5, 6 or 7 specifies that the input chroma samples are left-shifted by 0, 1, 2, 3, 4, 5, 6, or 7 bits, respectively, in the colorization process, for the sake of colorization.

4. Operation at the decoder

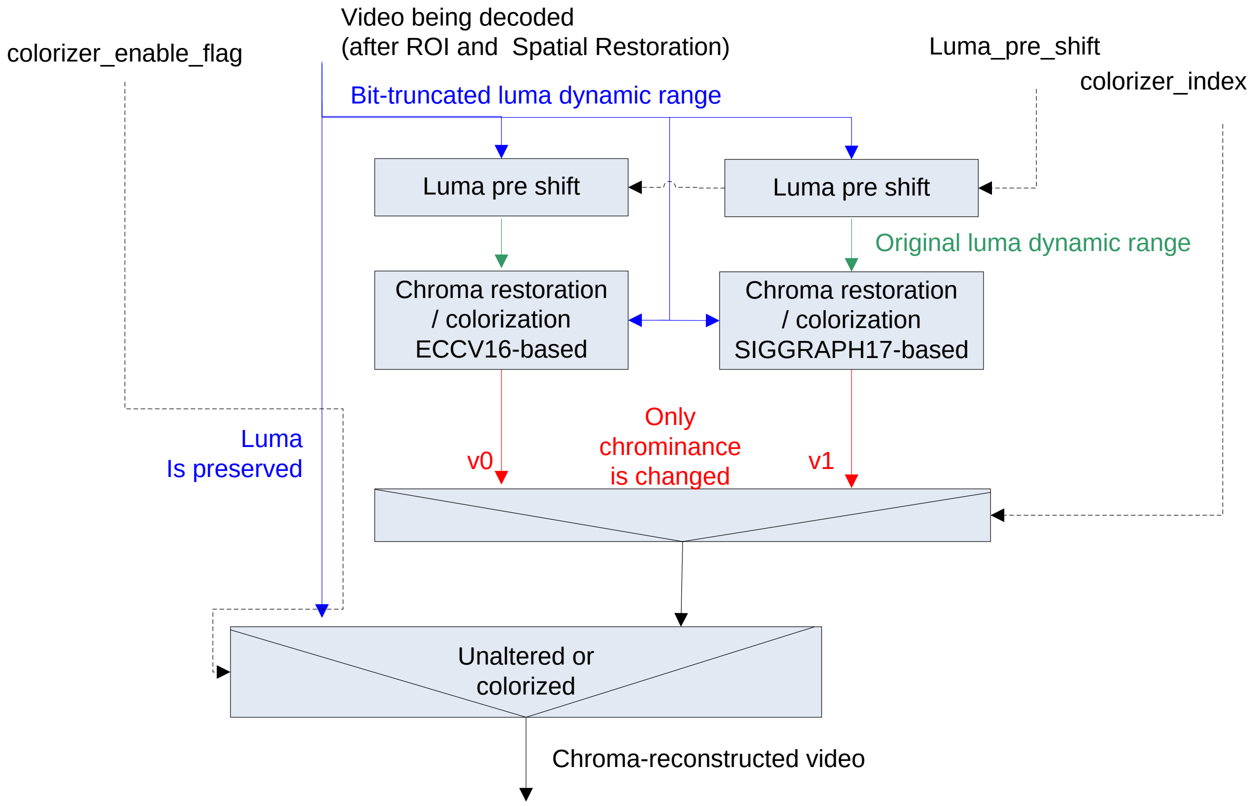

Fig.4. Operation of the proposed tool in the decoder.

When the tool is enabled (colorizer_enable_flag), the transmitted image is decolorized (contains luminance only).
The chrominance data is reconstructed with the use of one of the neural networks: ECCV16-based [2] or SIGGRAPH17-based [3].
The operation is as follows:

- First, the decoded luma component is left-shifted, based on signalled information: colorizer_luma_pre_shift value. This is done to restore the original luminance dynamic range, which might have been dimmed by BitTruncation tool at the encoder.
- No color conversion is performed – luma component is directly used (instead of L component from CIE Lab component as previously in [1])
- Before processing with neural network, the image is scaled to size 256x256 with the use of bilinear interpolation (already defined in WD spec).
- The luma component is fed the appropriate colorization neural network, based on colorized_index value
  - Value 0 means usage of ECCV16-based [2] network.
  - Value 1 means usage of SIGGRAPH17-based [3] network.
- The result of the network are CbCr components in YCbCr colorspace (instead of ab in CIE Lab colorpace as previously in [1])

- Finally, the image is back to the original resolution with the use bilinear interpolation (already defined in WD spec).
- Importantly Y component of the resulting image is copied from the original image. Effectively, only chrominance Cb and Cr channels are modified.

5. Neural Network models adaptation to YCbCr colorspace

The originally used ECCV17 [2] and SIGGRAPH18 [3] networks have been adapted:

- to work on luma *Y* component (from *Y Cb Cr*) component instead on *L* component (from CIE Lab),
- to output *Cb Cr* components (from *Y Cb Cr* color space) instead on *a b* (from CIE Lab).

In order to achieve this:

- Additional input and output layers have been incorporated in their respective architectures (to cover different color space)
- The additional layers have been initialized using training based on randomly generated *YCbCr* vs *Lab* samples.
- Additional retraining was performed of the overall model was performed but practically resulted in no improvement.

The whole adaptation was performed without any additional datasets.
No VCM datasets have been used. In particular ImageNet [5] dataset has been used.
As stated on https://www.image-net.org/ website:

> *© 2020 Stanford Vision Lab, Stanford University, Princeton University imagenet.help.desk@gmail.com Copyright infringement*
>
> *[…] The data is available for free to researchers for non-commercial use.*

To clarify licensing information an email has been sent to imagenet.help.desk@gmail.com. No response has been received yet.

In parallel to this, an attempt to analyze derivative-works legal situation has been made.
Apparently, "derivative status" of works (e.g. weights of ANN trained based on a dataset) depend on legal system which varies among countries. Application of particular legal system depends on the place of the act.
E.g. according to information found [6], Japan explicitly made this (usage of datasets for training networks without the stigma of derivative works) universally legal a couple years ago, allegedly in hopes of attracting ML talent.

As found on [7]:

> *Even when the copyright holder of the copyrighted work is a foreign company (or a foreign national) or when the user of the copyrighted work is a foreign company, Article 30-4 of Japan's Copyright Act will apply and be lawful as long as the place of the act of use of such copyrighted work is located in Japan.*

Please note that the information provided above is for informational purposes only and should not be construed as legal advice. For advice on specific legal matters, it is recommended to consult with a qualified legal professional.

6. Experimental results

The experiments have been performed in accordance to CTC.
The results have been collected to default excel sheet.
The summary of the results is presented in Tables 1, 2 and 3.
It can be notices that the gains are higher than in the previous version of the tool [1], probably due to:

- Lack of redundant color-space conversion
- Different Spatial Resampling settings (higher quality of the image to be colorized)
- Slightly better trained networks.

Table 1. Experimental results for Random Access: inner and end-to-end scenario. (FIRST FRAME DECISION)

RA_inner	Object Detection	BD-Rate [%] mAP	EncT	DecT
SFU-HW	Class A	0.00%	100.18%	99.68%
	Class B	0.00%	100.15%	100.41%
	Class C	0.00%	99.58%	99.58%
	Class D	0.00%	100.39%	99.83%
	Average	0.00%	100.04%	99.88%
	Class O	1.76%	100.43%	99.57%

RA_e2e	Object Detection	BD-Rate [%] mAP	EncT	DecT
SFU-HW	Class A	0.00%	99.51%	100.06%
	Class B	0.00%	100.03%	99.61%
	Class C	0.00%	100.42%	100.15%
	Class D	0.00%	100.35%	100.12%
	Average	0.00%	100.22%	100.00%
	Class O	7.26%	100.09%	99.87%

RA_inner	Object Tracking	BD-Rate [%] MOTA	EncT	DecT
TVD	TVD-01-1	0.00%	100.36%	100.04%
	TVD-01-2	0.00%	99.99%	100.07%
	TVD-01-3	0.00%	99.63%	100.02%
	TVD-02-1	0.00%	100.12%	100.17%
	TVD-03-1	-18.59%	100.43%	100.59%
	TVD-03-2	-30.15%	100.73%	100.52%
	TVD-03-3	-70.57%	100.83%	100.76%
	Average	-17.04%	100.30%	100.31%

RA_e2e	Object Tracking	BD-Rate [%] MOTA	EncT	DecT
TVD	TVD-01-1	0.00%	99.89%	99.66%
	TVD-01-2	0.00%	99.48%	100.04%
	TVD-01-3	0.00%	99.92%	99.77%
	TVD-02-1	0.00%	99.54%	99.39%
	TVD-03-1	-13.31%	101.00%	100.37%
	TVD-03-2	-27.58%	101.02%	101.44%
	TVD-03-3	-54.03%	100.25%	101.00%
	Average	-13.56%	100.15%	100.24%

Table 2. Experimental results for Low Delay: inner and end-to-end scenario. (FIRST FRAME DECISION)

LD_inner	Object Detection	BD-Rate [%] mAP	EncT	DecT
SFU-HW	Class A	0.00%	99.71%	100.11%
	Class B	0.00%	99.90%	99.57%
	Class C	0.00%	100.29%	100.26%
	Class D	0.00%	100.73%	99.71%
	Average	0.00%	100.29%	99.89%
	Class O	3.57%	100.41%	99.97%

LD_e2e	Object Detection	BD-Rate [%] mAP	EncT	DecT
SFU-HW	Class A	0.00%	100.19%	100.57%
	Class B	0.00%	100.17%	100.03%
	Class C	0.00%	100.11%	100.02%
	Class D	0.00%	100.75%	99.81%
	Average	0.00%	100.34%	100.00%
	Class O	1.28%	100.75%	99.77%

LD_inner	Object Tracking	BD-Rate [%] MOTA	EncT	DecT
TVD	TVD-01-1	0.00%	100.05%	100.11%
	TVD-01-2	0.00%	99.38%	99.45%
	TVD-01-3	0.00%	100.05%	100.23%
	TVD-02-1	0.00%	100.60%	99.82%
	TVD-03-1	-0.01%	100.89%	100.86%
	TVD-03-2	-26.71%	100.28%	102.10%
	TVD-03-3	-57.88%	100.97%	100.72%
	Average	-12.09%	100.32%	100.47%

LD_e2e	Object Tracking	BD-Rate [%] MOTA	EncT	DecT
TVD	TVD-01-1	0.00%	99.18%	99.92%
	TVD-01-2	0.00%	100.55%	99.55%
	TVD-01-3	0.00%	100.47%	100.06%
	TVD-02-1	0.00%	100.16%	99.79%
	TVD-03-1	-6.55%	101.20%	101.34%
	TVD-03-2	-17.83%	101.12%	101.94%
	TVD-03-3	-37.63%	101.48%	101.14%
	Average	-8.86%	100.59%	100.53%

Table 3. Experimental results for All Intra: inner and end-to-end scenario. (FIRST FRAME DECISION)

AI_inner	Object Detection	BD-Rate [%] mAP	EncT	DecT
SFU-HW	Class A	0.00%	100.13%	99.67%
	Class B	-0.01%	100.37%	99.58%
	Class C	-0.01%	100.68%	100.01%
	Class D	0.00%	100.31%	100.15%
	Average	0.00%	100.43%	99.92%
	Class O	1.62%	100.91%	99.51%

AI_e2e	Object Detection	BD-Rate [%] mAP	EncT	DecT
SFU-HW	Class A	0.00%	100.68%	99.40%
	Class B	0.00%	99.75%	99.91%
	Class C	0.00%	100.33%	100.06%
	Class D	0.00%	100.15%	99.79%
	Average	0.00%	100.15%	99.88%
	Class O	#	100.54%	100.11%

AI_inner	Object Tracking	BD-Rate [%] MOTA	EncT	DecT
TVD	TVD-01-1	0.00%	100.05%	99.45%
	TVD-01-2	0.00%	99.64%	99.89%
	TVD-01-3	0.00%	99.71%	100.04%
	TVD-02-1	0.00%	98.98%	100.72%
	TVD-03-1	-26.75%	101.58%	100.42%
	TVD-03-2	-30.70%	102.50%	101.12%
	TVD-03-3	-48.98%	102.67%	101.13%
	Average	-15.21%	100.72%	100.39%

AI_e2e	Object Tracking	BD-Rate [%] MOTA	EncT	DecT
TVD	TVD-01-1	0.00%	100.57%	99.97%
	TVD-01-2	0.00%	99.61%	99.69%
	TVD-01-3	0.00%	100.28%	100.34%
	TVD-02-1	0.00%	100.20%	100.05%
	TVD-03-1	-42.77%	107.39%	556.15%
	TVD-03-2	-37.34%	107.33%	231.87%
	TVD-03-3	-77.89%	108.50%	566.71%
	Average	-22.57%	103.34%	184.63%

6. Supplemental version – Dynamic decisions

During the 149th MPEG meeting it has been requested to assess situation in which the decision are not made against requirements of specific encoding scenarios. The following has been assumed:

- Random Access – decision made on the basis of Intra Periods
- All Intra – on the basis all every subsequent frame.
- Low Delay – on the basis of the first frame in the sequence as this does not break any constraints.

This has been addresses as follows:

1. An supplemental version of the proposal has been prepared.
2. The software has been used to demonstrate decision made by the encoder.
   Please see: “Dynamic_decisions_per_GOP.xlsx” and examples in Table 4.
   It can be seen that the decisions are consistent in time, e.g. in a very few points the decision differ from decision made on the basis of the first frame in the sequence.
3. The entire encoding experiment has been launched in such mode (the results are almost done and soon to be provided)

Table 4. Examples of dynamic decision making:

RA_e2e	TVD	0	TVD-01_1	x	x	x	x	x	x	x	x		
RA_e2e	TVD	1	TVD-01_2	x	x	x	x	x	x	x	x		
RA_e2e	TVD	2	TVD-01_3	x	x	x	x	x	x	x	x		
RA_e2e	TVD	3	TVD-02_1	x	x	x	x	x	x	x	x	x	...
RA_e2e	TVD	4	TVD-03_1	V	V	V	V	V	V	V	V		
RA_e2e	TVD	5	TVD-03_2	V	V	V	V	V	V	V	V		
RA_e2e	TVD	6	TVD-03_3	V	V	V	V	V	V	V	V		

AI_e2e	SFU	0	Traffic	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	...
AI_e2e	SFU	1	ParkScene	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	...
AI_e2e	SFU	2	BasketballDrive	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	...
AI_e2e	SFU	3	BQTerrace	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	...
AI_e2e	SFU	4	BasketballDrill	x	V	V	V	x	V	x	x	x	x	X	x	x	x	x	x	x	x	...
AI_e2e	SFU	5	BQMall	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	...
AI_e2e	SFU	6	PartyScene	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	...
AI_e2e	SFU	7	RaceHorsesC	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	...
AI_e2e	SFU	8	BasketballPass	x	x	x	x	x	x	x	x	X	x	x	x	x	x	x	x	x	x	...
AI_e2e	SFU	9	BQSquare	x	x	x	x	x	x	x	x	X	x	x	V	x	x	x	x	x	V	...
AI_e2e	SFU	10	BlowingBubbles	x	x	x	x	x	x	x	x	X	x	x	x	x	x	x	x	x	x	...
AI_e2e	SFU	11	RaceHorsesD	x	x	x	x	x	x	x	x	X	x	x	x	x	x	x	x	x	x	...
AI_e2e	SFU	12	Kimono	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	V	...
AI_e2e	SFU	13	Cactus	x	x	x	x	x	x	x	x	X	x	x	x	x	x	x	x	x	x	...

Table 5. Experimental results for Random Access: inner and end-to-end scenario.
.Left:FIRST FRAME DECISION,   Right: DYNAMIC DECISION

RA_inner	Object Detection	BD-Rate [%] mAP	BD-Rate [%] mAP
SFU-HW	Class A	0.00%	0.02%
	Class B	0.00%	0.02%
	Class C	0.00%	0.01%
	Class D	0.00%	0.01%
	Average	0.00%	0.01%
	Class O	1.76%	1.72%

RA_e2e	Object Detection	BD-Rate [%] mAP	BD-Rate [%] mAP
SFU-HW	Class A	0.00%	0.02%
	Class B	0.00%	0.01%
	Class C	0.00%	0.01%
	Class D	0.00%	0.01%
	Average	0.00%	0.01%
	Class O	7.26%	7.40%

RA_inner	Object Tracking	BD-Rate [%] MOTA	BD-Rate [%] MOTA
TVD	TVD-01-1	0.00%	0.01%
	TVD-01-2	0.00%	0.02%
	TVD-01-3	0.00%	0.01%
	TVD-02-1	0.00%	0.01%
	TVD-03-1	-18.59%	-18.49%
	TVD-03-2	-30.15%	-29.73%
	TVD-03-3	-70.57%	-70.50%
	Average	-17.04%	-16.95%

RA_e2e	Object Tracking	BD-Rate [%] MOTA	BD-Rate [%] MOTA
TVD	TVD-01-1	0.00%	0.02%
	TVD-01-2	0.00%	0.03%
	TVD-01-3	0.00%	0.01%
	TVD-02-1	0.00%	0.01%
	TVD-03-1	-13.31%	-13.01%
	TVD-03-2	-27.58%	-27.36%
	TVD-03-3	-54.03%	-54.03%
	Average	-13.56%	-13.48%

Table 6. Experimental results for Low Delay: inner and end-to-end scenario .Left:FIRST FRAME DECISION, Right: DYNAMIC DECISION

LD_inner	Object Detection	BD-Rate [%] mAP	BD-Rate [%] mAP
SFU-HW	Class A	0.00%	0.01%
	Class B	0.00%	0.01%
	Class C	0.00%	0.01%
	Class D	0.00%	0.01%
	Average	0.00%	0.01%
	Class O	3.57%	3.81%

LD_e2e	Object Detection	BD-Rate [%] mAP	BD-Rate [%] mAP
SFU-HW	Class A	0.00%	0.12%
	Class B	0.00%	0.17%
	Class C	0.00%	0.14%
	Class D	0.00%	0.20%
	Average	0.00%	0.17%
	Class O	1.28%	1.95%

LD_inner	Object Tracking	BD-Rate [%] MOTA	BD-Rate [%] MOTA
TVD	TVD-01-1	0.00%	0.01%
	TVD-01-2	0.00%	0.01%
	TVD-01-3	0.00%	0.01%
	TVD-02-1	0.00%	0.01%
	TVD-03-1	-0.01%	-0.31%
	TVD-03-2	-26.71%	-26.78%
	TVD-03-3	-57.88%	-57.92%
	Average	-12.09%	-12.14%

LD_e2e	Object Tracking	BD-Rate [%] MOTA	BD-Rate [%] MOTA
TVD	TVD-01-1	0.00%	0.29%
	TVD-01-2	0.00%	0.44%
	TVD-01-3	0.00%	0.25%
	TVD-02-1	0.00%	0.28%
	TVD-03-1	-6.55%	-6.15%
	TVD-03-2	-17.83%	-17.38%
	TVD-03-3	-37.63%	-37.20%
	Average	-8.86%	-8.50%

Table 7. Experimental results for All Intra: inner and end-to-end scenario.
.Left:FIRST FRAME DECISION,   Right: DYNAMIC DECISION

AI_inner	Object Detection	BD-Rate [%] mAP	BD-Rate [%] mAP	AI_e2e	Object Detection	BD-Rate [%] mAP	BD-Rate [%] mAP
SFU-HW	Class A	0.00%	0.01%	SFU-HW	Class A	0.00%	0.01%
	Class B	-0.01%	0.03%		Class B	0.00%	0.02%
	Class C	-0.01%	0.02%		Class C	0.00%	1.25%
	Class D	0.00%	0.27%		Class D	0.00%	0.40%
	Average	0.00%	0.10%		Average	0.00%	0.55%
	Class O	1.62%	1.57%		Class O	#	#

AI_inner	Object Tracking	BD-Rate [%] MOTA	BD-Rate [%] MOTA	AI_e2e	Object Tracking	BD-Rate [%] MOTA	BD-Rate [%] MOTA
TVD	TVD-01-1	0.00%	0.10%	TVD	TVD-01-1	0.00%	0.04%
	TVD-01-2	0.00%	0.10%		TVD-01-2	0.00%	0.04%
	TVD-01-3	0.00%	0.10%		TVD-01-3	0.00%	0.03%
	TVD-02-1	0.00%	0.06%		TVD-02-1	0.00%	0.03%
	TVD-03-1	-26.75%	-26.08%		TVD-03-1	-42.77%	-40.74%
	TVD-03-2	-30.70%	-30.29%		TVD-03-2	-37.34%	-37.28%
	TVD-03-3	-48.98%	-48.73%		TVD-03-3	-77.89%	-74.95%
	Average	-15.21%	-14.96%		Average	-22.57%	-21.83%

The results clearly show that the performance of the tool is practically the same in both operation modes:

- decision made based on the first frame
- decision made dynamically.

7. Supplemental version – Dynamic decisions with numerical stabilization

During additional cross-check by Ayoung Kim from ETRI it has been noticed that the tool produces different results when different platforms are used. We would like to express gratitude for identifying the problem.

The issue has been fixed by usage of quantized convolution scheme as already practiced in Temporal Resampling tool – by usage of IntConv2d operation instead of Conv2d.
Therefore the structure of the network is the same, but computational back-end now performs computations in quantized domain.

After the change the code has been checked on BasketballDrill and Kendo sequences and it produces consistent results among vastly different hardware platforms:

- Intel i7-5820K, i7-12700K, i7-13700K
- AMD Threadripper 1950X, 2950X, 3970X, R9 3900XT, R9 5950X, R9 7900X, R9 7950X

Additionally, to speed-up computations and to ensure reproducibility of the encoder bitstream, support for descriptors has been added, similar to ROI/Spatial ones.

Also, although the code is based on VCMRS v0.11.1, the adopted bit depth shift clean-up (m71663) has been included so that now syntax elements (luma shift) are no longer duplicated.

Summary of the changes:

- Inclusion of adopted m71663 (bit depth shift cleanup) which solved problem of additional sending of luma shift.
- Support for descriptors (similar to ROI and Spatial) have been added to allow encoder decision reproducibility. Currently v0.11.1 descriptor philosophy is applied (separate descriptors for different temporal resampling schemes). Later this will be aligned with the current v0.12RC descriptors scheme during integration phase.
- Conv2D has been replaced with IntConv2D just like in Temporal Resampling.
- PyTorch has been switched to deterministic mode.
- Some function, e.g. Pearson correlation has been rewritten for increased numerical stability and reproducibility.
- Together this solved the problem of numerical stability.
- Redundant code removal and clean-up.

The BD-Rate results of the code (m70724_chroma_reconstruction_dynamic_stable.xlsm) differ only by about +/-0.5p.p. as compared to dynamic variant (m70724_chroma_reconstruction_dynamic.xlsm).

8. Summary and recommendations

Summary:

- Improved proposed colorization (chrominance reconstruction) tool
    - Usage of Y Cb Cr colorspace instead of CIE Lab.
- Employs already defined neural-network-based processing framework in the decoder.
- Automatic control mechanism at the encoder based on similarity comparison between reconstructed chrominance and the original chrominance.
- Simple syntax and signalization.
- Gains: up to 22.57% of average gain (in Random Access end-to-end scenario, for Object Tracking task), when decisions done once per sequence, based on the first frame
- Practically the same performance when the decision is performed dynamically.

Recommendations:

- Adopt the proposed tool colorization tool to the VCM.

9. Acknowledgement

Thanks to Huifen Wang from China Telecom for performing cross-check of the baseline version of the proposal.
Thanks to Motong Xu and Roman Chernyak from Tencent for performing cross-check of two supplemental versions of the proposal.
Thanks to Ayoung Kim from ETRI for performing partial cross-check of the final supplemental version of the proposal.

10. References

[1] Olgierd Stankiewicz, Sławomir Różek, Mateusz Lorkiewicz, Tomasz Grajek, Sławomir Maćkowiak, Maciej Wawrzyniak, Jakub Stankowski, Marek Domański "[VCM] Neural-network-based chroma reconstruction" ISO/IEC JTC 1/SC 29/WG 4 m70249, 2024, Kemer, October 2024.

[2] Richard Zhang, Phillip Isola, Alexei A. Efros, "Colorful Image Colorization", Proceedings of ECCV, 2017

[3] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S. Lin, Tianhe Yu, Alexei A. Efros, "Real-Time User-Guided Image Colorization with Learned Deep Priors" ACM Transactions on Graphics (TOG), vol9, no 4. 2017.

[4] https://github.com/richzhang/colorization

[5] https://www.image-net.org/

[6] https://www.reddit.com/r/MachineLearning/comments/id4394/d_is_it_legal_to_use_models_pretrained_on/?rdt=36316

[7] https://storialaw.jp/en/service/bigdata/ai-14/14-2

Patent rights declaration(s)

Poznan University of Technology may have patent claims relating to the technology described in this contribution and, conditioned on reciprocity, is prepared to grant licenses under reasonable and non-discriminatory terms for such patent claims required to implement the resulting ITU-T Recommendation | ISO/IEC International Standard (per box 2 of the ITU-T/ITU-R/ISO/IEC patent statement and licensing declaration form).