



(51) International Patent Classification:

G11B 27/031 (2006.01) G06T 5/00 (2006.01)

(21) International Application Number:

PCT/US2023/022250

(22) International Filing Date:

15 May 2023 (15.05.2023)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

63/342,306 16 May 2022 (16.05.2022) US
22182506.0 01 July 2022 (01.07.2022) EP

(71) Applicant: **DOLBY LABORATORIES LICENSING CORPORATION** [US/US]; 1275 Market Street, San Francisco, California 94103 (US).

(72) Inventors: **MUSUNURI, Sri Harsha**; c/o Dolby Laboratories, Inc., 1275 Market Street, San Francisco, California 94103 (US). **ROTTI, Shruthi Suresh**; c/o Dolby Laboratories, Inc., 1275 Market Street, San Francisco, California 94103 (US). **CHOUDHURY, Anustup Kumar Atanu**; c/o Dolby Laboratories, Inc., 1275 Market Street, San Francisco, California 94103 (US).

(74) Agent: **KONSTANTINIDES, Konstantinos** et al.; Dolby Laboratories, Inc., Intellectual Property Group, 1275 Market Street, San Francisco, California 94103 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT,

LU, LV, MC, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— with international search report (Art. 21(3))

(54) Title: TRIM PASS METADATA PREDICTION IN VIDEO SEQUENCES USING NEURAL NETWORKS

(57) Abstract: Methods and systems for generating trim-pass metadata for high dynamic range (HDR) video are described. The trim-pass prediction pipeline includes a feature extraction network followed by a fully connected network which maps extracted features to trim-pass values. In a first architecture, the feature extraction network is based on four cascaded convolutional networks. In a second architecture, the feature extraction network is based on a modified MobileNetV3 neural network. In both architectures, the fully connected network is formed by a set of three linear networks, each set customized to best match its corresponding feature extraction network.

TRIM PASS METADATA PREDICTION IN VIDEO SEQUENCES USING NEURAL NETWORKS

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of priority from U.S. Provisional Patent Application Ser. No. 63/342,306, filed on May 16, 2022, and European Patent Application No. 22 182 506.0, filed on July 1, 2022, each of which is incorporated by reference in its entirety.

TECHNOLOGY

[0002] The present invention relates generally to images. More particularly, an embodiment of the present invention relates to techniques for predicting trim pass metadata in video sequences using neural networks.

BACKGROUND

[0003] As used herein, the term 'dynamic range' (DR) may relate to a capability of the human visual system (HVS) to perceive a range of intensity (e.g., luminance, luma) in an image, e.g., from darkest grays (blacks) to brightest whites (highlights). In this sense, DR relates to a 'scene-referred' intensity. DR may also relate to the ability of a display device to adequately or approximately render an intensity range of a particular breadth. In this sense, DR relates to a 'display-referred' intensity. Unless a particular sense is explicitly specified to have particular significance at any point in the description herein, it should be inferred that the term may be used in either sense, e.g., interchangeably.

[0004] As used herein, the term high dynamic range (HDR) relates to a DR breadth that spans some 14-15 orders of magnitude of the human visual system (HVS). In practice, the DR over which a human may simultaneously perceive an extensive breadth in intensity range may be somewhat truncated, in relation to HDR. As used herein, the terms enhanced dynamic range (EDR) or visual dynamic range (VDR) may individually or interchangeably relate to the DR that is perceivable within a scene or image by a human visual system (HVS) that includes eye movements, allowing for some light adaptation changes across the scene or image.

[0005] In practice, images comprise one or more color components (e.g., luma Y and chroma Cb and Cr) wherein each color component is represented by a precision of n -bits per pixel (e.g., $n = 8$). For example, using gamma luminance coding, images where $n \leq 8$ (e.g.,

color 24-bit JPEG images) are considered images of standard dynamic range, while images where $n \geq 10$ may be considered images of enhanced dynamic range. EDR and HDR images may also be stored and distributed using high-precision (e.g., 16-bit) floating-point formats, such as the OpenEXR file format developed by Industrial Light and Magic.

[0006] Most consumer desktop displays currently support luminance of 200 to 300 cd/m² or nits. Most consumer HDTVs range from 300 to 500 nits with new models reaching 1000 nits (cd/m²). Such conventional displays thus typify a lower dynamic range (LDR), also referred to as a standard dynamic range (SDR), in relation to HDR or EDR. As the availability of HDR content grows due to advances in both capture equipment (e.g., cameras) and HDR displays (e.g., the PRM-4200 professional reference monitor from Dolby Laboratories), HDR content may be color graded and displayed on HDR displays that support higher dynamic ranges (e.g., from 1,000 nits to 5,000 nits or more). In general, without limitation, the methods of the present disclosure relate to any dynamic range higher than SDR.

[0007] As used herein, the term “display management” refers to processes that are performed on a receiver to render a picture for a target display. For example, and without limitation, such processes may include tone-mapping, gamut-mapping, color management, frame-rate conversion, and the like.

[0008] As used herein, the term “trim-pass” refers to a video post-production process where a colorist or a creative in charge of the content goes through a master grade of the content, shot by shot, and adjusts the lift-gamma-gain primaries and/or other color parameters to create a desired color or effect. Parameters related to this process (say, lift, gain, and gamma values) may also be embedded as trim-pass metadata or “trims” within the video content to be used later on as part of the display management process.

[0009] The creation and playback of high dynamic range (HDR) content is now becoming widespread as HDR technology offers more realistic and lifelike images than earlier formats. However, the broadcast infrastructure may not support the generation and transmission of custom trims when converting HDR content into SDR content for legacy displays. To improve existing coding schemes, as appreciated by the inventors here, improved techniques for automatically generating trim-pass metadata are developed.

[00010] US 2021/076042 A1 discloses a method for generating metadata for use by a video decoder for displaying video content encoded by a video encoder that includes accessing a target tone mapping curve; accessing a decoder tone curve corresponding to a tone curve used by the video decoder for tone mapping the video content; generating a plurality of parameters of a trim-pass function used by the video decoder to apply after applying the decoder tone

curve to the video content, wherein the parameters of the trim-pass function are generated to approximate the target tone curve with the combination of the trim-pass function and the decoder tone curve, and generating the metadata for use by the video decoder, including said plurality of parameters of the trim-pass function.

[00011] US 2021/350512 A1 discloses a method of automatic display management generation for gaming or SDR+ contents. Different candidate image data feature types are evaluated to identify one or more specific image data feature types to be used in training a prediction model for optimizing one or more image metadata parameters. A plurality of image data features of the one or more selected image data feature types is extracted from one or more images. The plurality of image data features of the one or more selected image data feature types is reduced into a plurality of significant image data features. A total number of image data features in the plurality of significant image data features is no larger than a total number of image data features in the plurality of image data features of the one or more selected image data feature types. The plurality of significant image data features is applied to training the prediction model for optimizing one or more image metadata parameters.

[00012] The approaches described in this section are approaches that could be pursued, but not necessarily approaches that have been previously conceived or pursued. Therefore, unless otherwise indicated, it should not be assumed that any of the approaches described in this section qualify as prior art merely by virtue of their inclusion in this section. Similarly, issues identified with respect to one or more approaches should not assume to have been recognized in any prior art on the basis of this section, unless otherwise indicated.

SUMMARY

[00013] The invention is defined by the independent claims. The dependent claims concern optional features of some embodiments of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

[00014] An embodiment of the present invention is illustrated by way of example, and not in way by limitation, in the figures of the accompanying drawings and in which like reference numerals refer to similar elements and in which:

[00015] FIG. 1A depicts an example process for predicting trim-pass metadata for HDR video according to an example embodiment of the present invention;

[00016] FIG. 1B depicts an example process for training the neural networks depicted in FIG. 1A;

[00017] FIG. 2 depicts a first example of a neural-network architecture for generating trim-pass metadata according to an example embodiment of the present invention; and

[00018] FIG. 3 depicts a second example of a neural-network architecture for generating trim-pass metadata according to another example embodiment of the present invention.

DESCRIPTION OF EXAMPLE EMBODIMENTS

[00019] Methods for trim-pass metadata prediction for video are described herein. In the following description, for the purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the present invention. It will be apparent, however, that the present invention may be practiced without these specific details. In other instances, well-known structures and devices are not described in exhaustive detail, in order to avoid unnecessarily occluding, obscuring, or obfuscating the present invention.

SUMMARY

[00020] Example embodiments described herein relate to methods for generating trim-pass metadata for video sequences. In an embodiment, a processor receives a picture in a video sequence. A feature extraction neural network extracts image features from the picture and then passes them to a fully connected neural network which maps the image features to output trim-pass metadata values for the input picture.

TRIM-PASS METADATA PREDICTION PIPELINE

[00021] In traditional display mapping (DM), the mapping algorithm applies a sigmoid like function (for examples, see Refs [1-2]) to map the input dynamic range to the dynamic range of the target display. Such mapping functions may be represented as piece-wise linear or non-linear polynomials characterized by anchor points, pivots, and other polynomial parameters generated using characteristics of the input source and the target display. For example, in Refs. [1-2] the mapping functions use anchor points based on luminance characteristics (e.g., the minimum, medium (average), and maximum luminance) of the input images and the display. However, other mapping functions may use different statistical data, such as luminance-variance or luminance-standard deviation values at a block level or for the whole image. For SDR images, the process may also be assisted by additional metadata which are either transmitted as part of the transmitted video or they are computed by the decoder or the

display. For example, when the content provider has both SDR and HDR versions of the source content, a source may use both versions to generate metadata (such as piece-wise linear approximations of forward or backward reshaping functions) to assist the decoder in converting incoming SDR images to HDR images.

[00022] As used herein, the term “L1 metadata” denotes minimum, medium, and maximum luminance values related to an input frame or image. L1 metadata may be computed by converting RGB data to a luma-chroma format (e.g., YCbCr) and then computing min, mid (average), and max values in the Y plane, or they can be computed directly in the RGB space. For example, in an embodiment, L1Min denotes the minimum of the PQ-encoded $\min(\text{RGB})$ values of the image, while taking into consideration an active area (e.g., by excluding gray or black bars, letterbox bars, and the like). $\min(\text{RGB})$ denotes the minimum of color component values {R, G, B} of a pixel. The values of L1Mid and L1Max may also be computed in a same fashion replacing the $\min()$ function with the $\text{average}()$ and $\text{max}()$ functions. For example, L1Mid denotes the average of the PQ-encoded $\text{max}(\text{RGB})$ values of the image, and L1Max denotes the maximum of the PQ-encoded $\text{max}(\text{RGB})$ values of the image. In some embodiments, L1 metadata may be normalized to be in [0, 1].

[00023] Considering the L1Min, L1Mid, and L1Max values of the original HDR metadata, as well as the maximum (peak) and minimum (black) luminance of the target display, denoted as Tmax and Tmin. Then, as described in Ref. [1-2], one may generate an intensity tone-mapping mapping curve mapping the intensity of the input image to the dynamic range of the target display. This may be considered to be the ideal, single-stage, tone-mapping curve, to be matched by using the reconstructed metadata.

[00024] As noted earlier, the term “trims” denotes tone-curve adjustments performed by a colorist to improve tone mapping operations. Trims are typically applied to the SDR range (e.g., 100 nits maximum luminance, 0.005 nits minimum luminance). These values are then interpolated linearly to the target luminance range depending only on the maximum luminance. These values modify the default tone curve and are present for every trim.

[00025] Information about the trims may be part of the HDR metadata and may be used to adjust the original tone-mapping curves (see Ref. [1-3]). For example, in Dolby Vision, trims may be passed as Level 2 (L2) or Level 8 (L8) metadata that includes Slope, Offset, and Power variables (collectively referred to as SOP parameters) representing Gain and Gamma values to adjust pixel values. For example, if Slope, Offset, and Power are in [-0.5, 0.5], then, given Lift, Gain and Gamma:

$$\text{Slope} = \max(-0.5, \min(0.5, \text{Gain} * (1 - \text{Lift}) - 1))$$

$$\text{Offset} = \max(-0.5, \min(0.5, \text{Gain} * \text{Lift})) \quad (1)$$

$$\text{Power} = \max(-0.5, \min(0.5, 1 / \text{Gamma} - 1))$$

[00026] In certain content creation scenarios, it may not be possible to employ a full range of color grading tools to derive SDR content from HDR content. Embodiments described herein propose using a neural network-based architecture to automatically generate such trim-pass metadata. The trim-pass metadata are configured to perform adjustments of a tone-mapping curve that is applied to an input picture when being displayed on a target display. While examples presented herein describe mostly how to predict Slope, Offset, and Power values, similar architectures can be applied to predict directly Lift, Gain, or Gamma, or other trims as those described in Ref. [3].

[00027] FIG. 1A depicts an example trim-prediction pipeline (100) for HDR images (102) according to an embodiment. As depicted in FIG. 1A, pipeline 100 includes the following modules or components:

- HDR input (102)
- a neural network for feature extraction (105)
- a fully-connected neural network to map the extracted features to trim-pass metadata (110)
- trim-pass metadata output (112) (e.g., predicted SOP data)

[00028] The network (100) takes a given frame (102) as input and passes it to a convolutional neural network (105) for feature extraction that identifies the high-level features of the image, which are then passed to a fully connected network (110) that provides the mapping between the features and the trim metadata (112). Each of the high-level features corresponds to an image feature type of a set of image feature types. Each image feature type is represented by a plurality of associated image features extracted from a corresponding set of training images and used for training the convolutional neural network (105) for feature extraction. This network can be used as a stand-alone component, or it can be integrated in video processing pipeline. For maintaining temporal consistency of the trim pass metadata, the network can also take multiple frames as input.

[00029] In an embodiment, the pipeline is formed out of neural-network (NN) blocks trained to work with HDR images coded using perceptual quantization (PQ) in the ICtCp color space as defined in Rec. BT. 2390, “*High dynamic range television for production and international programme exchange*.” The ICtCp space is preferred due to being a perceptually uniform color space, thus improving spatio-temporal consistency.

[00030] FIG. 1B depict an example pipeline (130) to train network 100. Compared to the system (100), this system also includes:

- Training input of true trim-pass metadata (132)
- An error/loss computation module (120), which computes an error function (say, mean square error (MSE) or mean absolute error (MAE)) between the training data (132) and the predicted values (112)
- A back propagation path (122) to train networks 105 and 110.

Trim-pass Prediction Architectures

[00031] Two different trim-pass prediction architectures (100) have been designed. They provide a balance between accuracy and speed. FIG. 2 depicts the first architecture according to an embodiment.

[00032] In an embodiment, a neural network may be defined as a set of 4-dimensional convolutions, each of which is followed by adding a constant bias value to all results. In some layers, the convolution is followed by clamping negative values to 0. The convolutions are defined by their size in pixels ($M \times N$), how many image channels (C) they operate on, and how many such kernels are in the filter bank (K). In that sense, each convolution can be described by the size of the filter bank $M \times N \times C \times K$. As an example, a filter bank of the size $3 \times 3 \times 1 \times 2$ is composed of 2 convolution kernels, each of which operates on one channel, and has a size of 3 pixels by 3 pixels. Input and Output sizes are denoted as Height x Width x Channels.

[00033] Some filter banks may also have a stride, meaning that some results of the convolution are discarded. A stride of 1 means every input pixel produces an output pixel. A stride of 2 means that only every second pixel in each dimension produces an output, and the like. Thus, a filter bank with a stride of 2 will produce an output with $(M/2) \times (N/2)$ pixels, where $M \times N$ is the input image size. If padding = 1, all inputs except the ones to fully connected kernels are padded so that setting the stride of 1 would produce an output with the

same number of pixels as the input. The output of each convolution bank feeds as an input into the next convolution layer.

[00034] As depicted in FIG. 2, in the first architecture, the feature extraction network (105) includes four convolutional networks configured as:

- CONV1: Input 540x960x1, 3x3x1x4, stride = 2, output: 270x480x4, bias, rectified linear unit (ReLU) activation
- CONV2: 3x3x4x8, stride = 2, bias, activation ReLU, output: 135x240x8
- CONV3: 7x7x8x16, stride = 5, bias, activation ReLU, output: 27x48x16
- CONV4: 27x48x16x3 (fully connected), stride = 1, bias, output: 1x1x3

[00035] Following the feature extraction network, the fully connected network (110) includes three linear networks configured as follows:

- Linear1: Input: 1x3, Output 1x6, Batch Norm, ReLU, DropOut
- Linear2: Input: 1x6, Output 1x6, Batch Norm, ReLU, DropOut
- Linear 3: Input: 1x6, Output 1x3

[00036] DropOut refers to using Dropout regularization, and BatchNorm refers to applying Batch Normalization (Ref. [5]) to increase training speed.

[00037] FIG. 3 depicts an example of the second architecture. As depicted in FIG. 3, in the second architecture, the feature extraction network (105) comprises a modified version of the MobileNet_V3 network, originally described in Ref. [4] for images with 1:1 aspect ratios, but now extended to operate on images of arbitrary aspect ratios. Table 1 (based on Table 2 in Ref.[4]) provides additional details for the operations in the modified MobileNetV3 architecture for an example 540x960x3 input. Compared to Ref.[4], after the second conv2d, 1x1, stage, the original pool, 7x7 stage is replaced by an average pool, 1x1 stage to generate the final output (e.g., 1x576), and the two subsequent conv2d, 1x1, non-batch normalization (NBN) stages are removed.

[00038] In Table 1, “bneck” denotes bottleneck and inverted residual blocks, Exp size denotes expansion block size, SE denotes if squeeze and excite is enabled, NL denotes the non-linear activation function, HS denotes the hard-swish activation function, RE denotes the ReLU activation function, and s denotes the stride. The output of a neural network stage matches the input of the subsequent neural network stage.

Table 1. Specification for MobileNetV3 (Small)

Input	Operator	Exp size	SE	NL	s
-------	----------	----------	----	----	-----

540x960x3	conv 2d, 3x3	-	-	HS	2
270x480x16	bneck, 3x3	16	Yes	RE	2
135x240x16	bneck, 3x3	72	-	RE	2
68x120x24	bneck, 3x3	88	-	RE	1
68x120x24	bneck, 5x5	96	Yes	HS	2
34x60x40	bneck, 5x5	240	Yes	HS	1
34x60x40	bneck, 5x5	240	Yes	HS	1
34x60x40	bneck, 5x5	120	Yes	HS	1
34x60x48	bneck, 5x5	144	Yes	HS	1
34x60x48	bneck, 5x5	288	Yes	HS	2
17x30x96	bneck, 5x5	576	Yes	HS	1
17x30x96	bneck, 5x5	576	Yes	HS	1
17x30x96	conv2d, 1x1	-	Yes	HS	1
17x30x576	average pool, 1x1	-	-	-	1
Output 1x1x576 to the Fully Connected Network					

[00039] Following network 105, the fully connected network (110) includes three linear networks configured as follows:

- Linear1: Input: 1x576, Output: 1x128, Batch Norm, ReLU, DropOut
- Linear2: Input: 1x128, Output: 1x64, Batch Norm, ReLU, DropOut
- Linear 3: Input: 1x64, Output: 1x3

[00040] As described in Ref. [4], MobileNetV3 is tuned to mobile phone CPUs for object detection and semantic segmentation (or dense pixel prediction). It uses a variety of innovative tools, including:

- Depth-wise convolutional filters
- Bottleneck and inverted residual blocks (bneck)
- Squeeze and excite (SE) blocks (Ref. [6]); and the
- Hard-Swish (HS) activation function, defined as

$$hard_swish[x] = x \frac{ReLU6(x+3)}{6}, \quad (2)$$

where ReLU6 denotes a known in the art activation function.

Fully Connected Network

[00041] The output of the Feature Extraction Module (105) is fed to the fully connected network to get the predicted trim values. Based on the architecture of the feature extraction module, the fully connected network could have a different input size and thus might need a different output size. This module learns the mapping between the high-level features extracted from the image to its associated Slope, Offset, and Power. As mentioned above, the same network design can be used to directly compute Lift, Gain and Gamma, or any other parameters related to trim pass metadata.

[00042] In terms of complexity, the first architecture is simpler and less computationally intensive than the MobileNet-based architecture; however, the MobileNet-based architecture is more accurate.

[00043] The second architecture can take directly three input channels (say, RGB or ICtCp); however, it would be straightforward to modify the first architecture to take into consideration chroma as well by allowing three input channels instead of just one.

Network Training

[00044] As depicted in FIG. 1B, in training, during the error computation (120), one may apply either an L1 loss (e.g., MAE) or an L2 loss (e.g., MSE). This loss is calculated between the predicted trim pass metadata values (112) and the ground truth trim pass metadata values (132). Alternatively, one may also use the tone curve as a loss function. Essentially, one can use the trim pass metadata values along with other metadata values (such as L1) and define a tone curve corresponding to that image. Then one can minimize either the L1 loss or L2 loss between the tone curves that are obtained using the predicted trim pass metadata and the ground truth trim pass metadata.

[00045] For example, let $f_{true}(i)$ denote a tone curve generated using L1 metadata for the input (102) and training trim-pass metadata (132), and let $f_{pred}(i)$ denote a tone curve generated using the same L1 metadata and predicted trim-pass metadata (112). Then, during training, one may want to minimize the mean-square error defined as:

$$Loss = \frac{1}{|i|} \sum_i (f_{true}(i) - f_{pred}(i))^2, \quad (3)$$

where $|i|$ denotes the cardinality of the i -values for which the MSE is computed.

Alternatively, one may apply an L1 metric, as in

$$Loss = \frac{1}{|i|} \sum_i |f_{true}(i) - f_{pred}(i)|. \quad (4)$$

References

Each of these references is include by reference in its entirety.

1. A. Ballestad and A. Kostin, U.S. Patent 8,593,480, “*Method and apparatus for image data transformation*”
2. J.A. Pytlarz and R. Atkins, U.S. Patent 10,600,166, “*Tone curve mapping for high dynamic range images.*”
3. “*Module 2.8 – The Dolby Vision metadata Trim Pass*,” “<https://learning.dolby.com/hc/en-us/articles/360056574431-Module-2-8-The-Dolby-Vision-Metadata-Trim-Pass->”, downloaded May 3, 2022.
4. A. Howard, et al., “*Searching for MobileNetV3*,” Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019, also arXiv:1905.02244v5, 20 Nov. 2019.
5. S. Ioffe, and C. Szegedy. “*Batch normalization: Accelerating deep network training by reducing internal covariate shift.*” International conference on machine learning. PMLR, 2015, also arXiv:1502.03167v3, 2 Mar. 2015.
6. J. Hu, et al, “*Squeeze-and-excitation networks*,” arXiv:1709.01507v4, 16 May 2019.

EXAMPLE COMPUTER SYSTEM IMPLEMENTATION

[00046] Embodiments of the present invention may be implemented with a computer system, systems configured in electronic circuitry and components, an integrated circuit (IC) device such as a microcontroller, a field programmable gate array (FPGA), or another configurable or programmable logic device (PLD), a discrete time or digital signal processor (DSP), an application specific IC (ASIC), and/or apparatus that includes one or more of such systems, devices or components. The computer and/or IC may perform, control, or execute instructions related to image transformations, such as those described herein. The computer and/or IC may compute any of a variety of parameters or values that relate to trim-pass metadata prediction processes described herein. The image and video embodiments may be implemented in hardware, software, firmware and various combinations thereof.

[00047] Certain implementations of the invention comprise computer processors which execute software instructions which cause the processors to perform a method of the invention. For example, one or more processors in a display, an encoder, a set top box, a transcoder or the like may implement methods related to trim-pass metadata prediction

processes as described above by executing software instructions in a program memory accessible to the processors. The invention may also be provided in the form of a program product. The program product may comprise any tangible and non-transitory medium which carries a set of computer-readable signals comprising instructions which, when executed by a data processor, cause the data processor to execute a method of the invention. Program products according to the invention may be in any of a wide variety of tangible forms. The program product may comprise, for example, physical media such as magnetic data storage media including floppy diskettes, hard disk drives, optical data storage media including CD ROMs, DVDs, electronic data storage media including ROMs, flash RAM, or the like. The computer-readable signals on the program product may optionally be compressed or encrypted.

[00048] Where a component (e.g. a software module, processor, assembly, device, circuit, etc.) is referred to above, unless otherwise indicated, reference to that component (including a reference to a "means") should be interpreted as including as equivalents of that component any component which performs the function of the described component (e.g., that is functionally equivalent), including components which are not structurally equivalent to the disclosed structure which performs the function in the illustrated example embodiments of the invention.

EQUIVALENTS, EXTENSIONS, ALTERNATIVES AND MISCELLANEOUS

[00049] Example embodiments that relate to trim-pass metadata prediction processes are thus described. In the foregoing specification, embodiments of the present invention have been described with reference to numerous specific details that may vary from implementation to implementation. Thus, the sole and exclusive indicator of what is the invention, and what is intended by the applicants to be the invention, is the set of claims that issue from this application, in the specific form in which such claims issue, including any subsequent correction. Any definitions expressly set forth herein for terms contained in such claims shall govern the meaning of such terms as used in the claims. Hence, no limitation, element, property, feature, advantage or attribute that is not expressly recited in a claim should limit the scope of such claim in any way. The specification and drawings are, accordingly, to be regarded in an illustrative rather than a restrictive sense.

CLAIMS

What is claimed is:

1. A method for generating trim-pass metadata of pictures in a video sequence, wherein the trim-pass metadata are configured to perform adjustments of a tone-mapping curve that is applied to an input picture when being displayed on a target display, the method comprising:
 - receiving the input picture;
 - providing a feature extraction network, the feature extraction network comprising a convolutional neural network for feature extraction that is trained to identify high-level image features of the input picture;
 - applying the feature extraction network to the input picture to generate the image features;
 - providing a fully connected network, the fully connected network comprising a plurality of cascaded linear neural networks that are trained to map the image features to output trim-pass metadata values for the input picture; and
 - applying the fully connected network to the image features to map the image features to the output trim-pass metadata values for the input picture.
2. The method of claim 1, wherein the input picture is a high-dynamic range (HDR) picture coded using PQ encoding in the ICtCp color space.
3. The method of claim 1 or claim 2, wherein the feature extraction network comprises four cascaded convolutional networks.
4. The method of claim 3, wherein the fully connected network comprises three cascaded linear networks.
5. The method of any of claims 1-4, wherein the feature extraction network comprises a modified MobileNetV3 neural network accepting inputs with non square aspect ratios.
6. The method of claim 5, wherein the fully connected network comprises three cascaded linear networks.

7. The method of any of claims 1-6, further comprising:
- receiving input training trim-pass parameters corresponding to the input picture;
 - applying an error loss unit to generate an error metric based on the input training trim-pass parameters and the output trim-pass metadata; and
 - training the feature extraction network and the fully connected network by minimizing the error metric.
8. The method of claim 7, wherein computing the error metric comprises computing a minimum absolute error or a mean square error between the input training trim-pass parameters and the output trim-pass metadata.
9. The method of claim 7, wherein computing the error metric comprises:
- generating a first tone-mapping function based at least on the input training trim-pass parameters;
 - generating a second tone-mapping function based at least on the output trim-pass metadata; and
 - computing the minimum absolute error or the mean square error between values of the first tone-mapping function and the second tone-mapping function.
10. An apparatus comprising a processor and configured to perform any one of the methods recited in claims 1-9.
11. A non-transitory computer-readable storage medium having stored thereon computer-executable instruction for executing a method with one or more processors in accordance with any one of the claims 1-9.

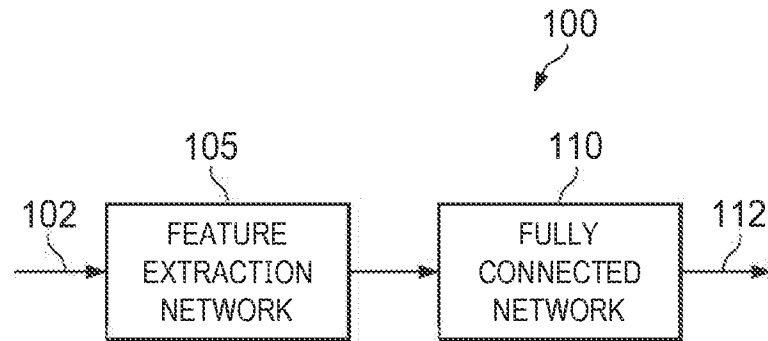


FIG. 1A

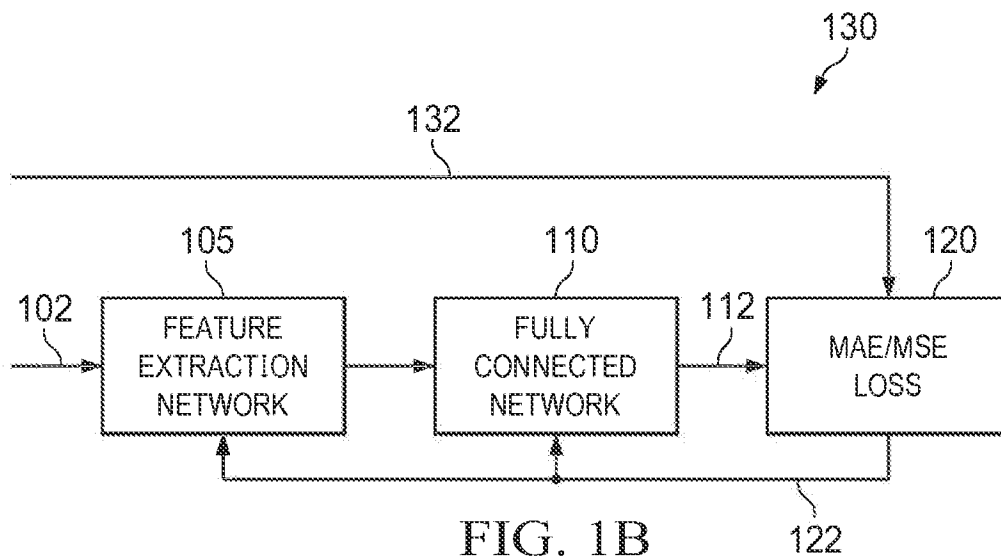


FIG. 1B

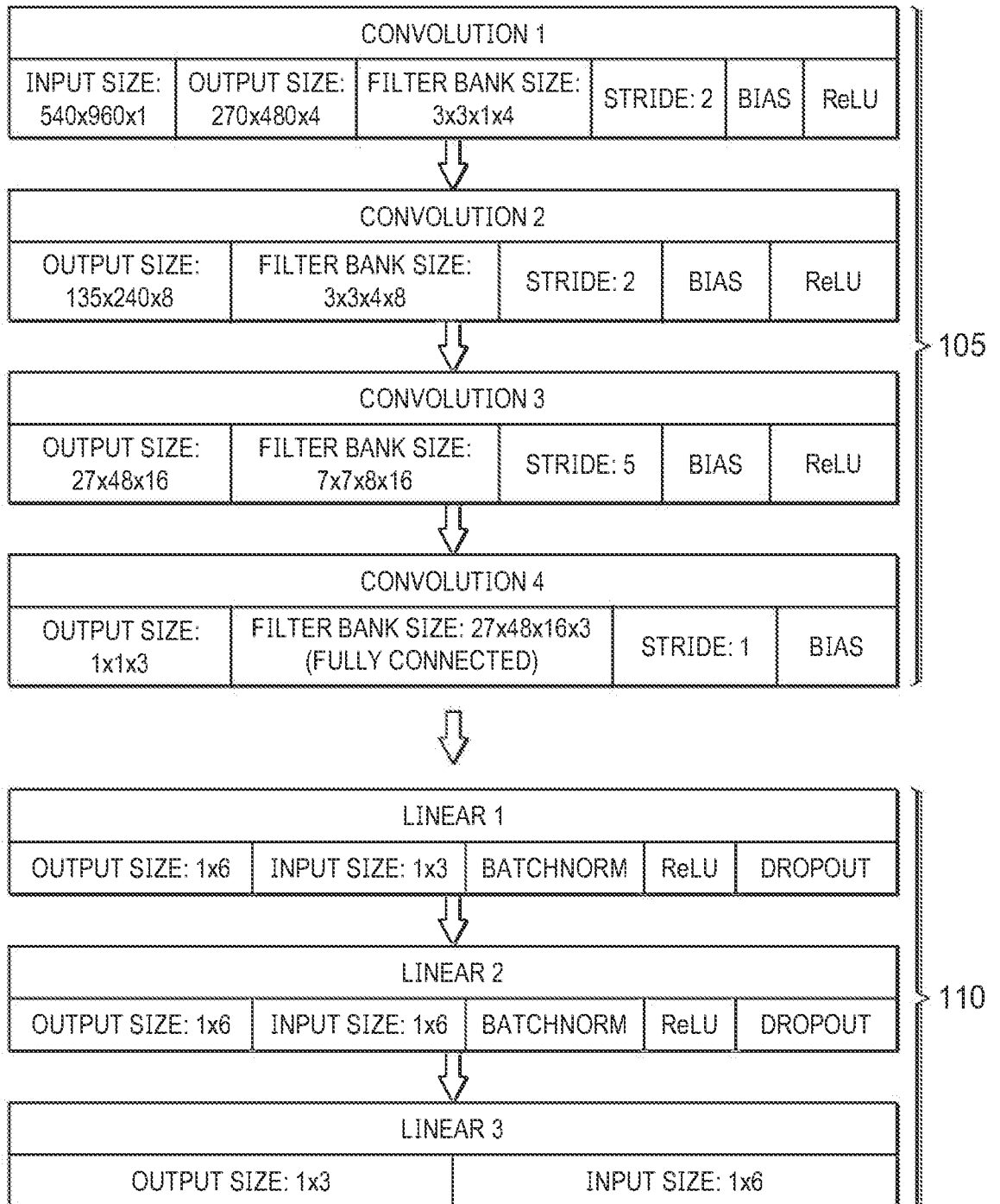


FIG. 2

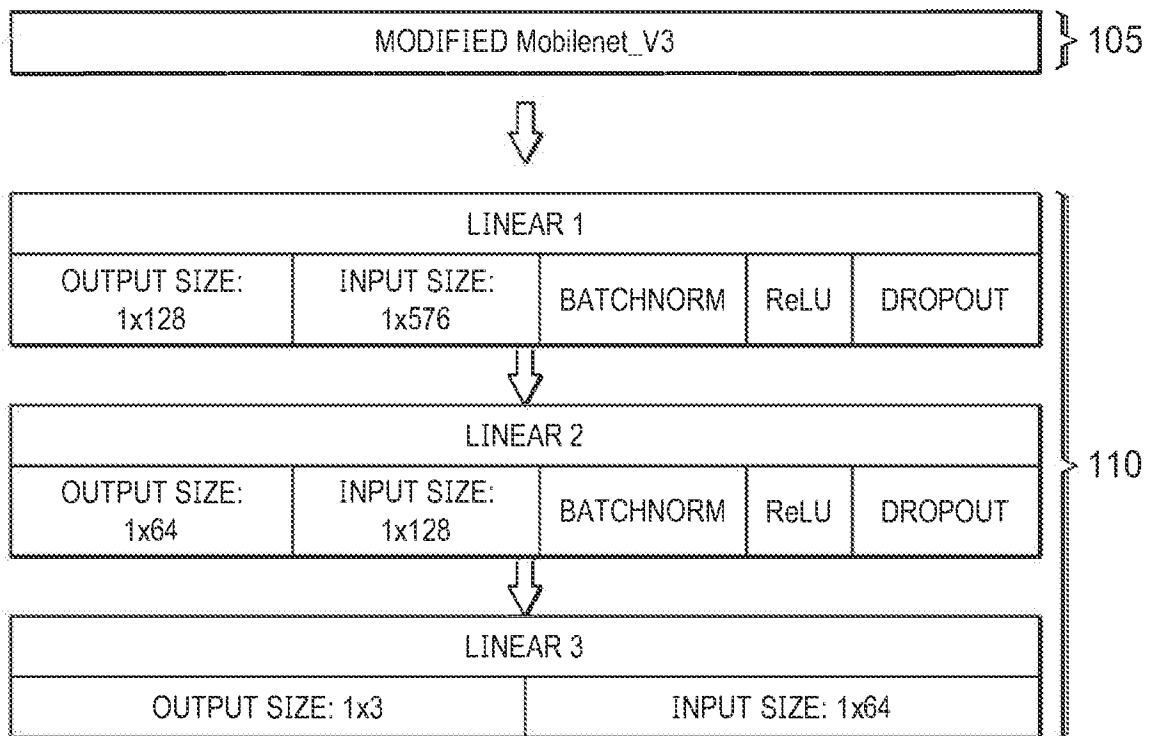


FIG. 3

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2023/022250

A. CLASSIFICATION OF SUBJECT MATTER
INV. G11B27/031 G06T5/00
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G11B G06T H04N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2021/076042 A1 (CHOUDHURY ANUSTUP KUMAR ATANU [US] ET AL) 11 March 2021 (2021-03-11) paragraphs [0018], [0048], [0049] -----	1-11
X	US 2021/350512 A1 (KADU HARSHAD [US] ET AL) 11 November 2021 (2021-11-11) paragraphs [0044] - [0051], [0193] -----	1-11

☐ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

17 August 2023

Date of mailing of the international search report

25/08/2023

Name and mailing address of the ISA/
European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Mourik, Piet

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2023/022250

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2021076042 A1	11-03-2021	CN 111095931 A	01-05-2020
		EP 3679716 A2	15-07-2020
		JP 7084984 B2	15-06-2022
		JP 2020533841 A	19-11-2020
		US 2021076042 A1	11-03-2021
		WO 2019050972 A2	14-03-2019

US 2021350512 A1	11-11-2021	CN 112714923 A	27-04-2021
		EP 3853810 A1	28-07-2021
		KR 20210046086 A	27-04-2021
		US 2021350512 A1	11-11-2021
		WO 2020061172 A1	26-03-2020
