(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(54) Title: METHOD AND APPARATUS FOR CONTEXT-BASED VIDEO QUALITY ASSESSMENT
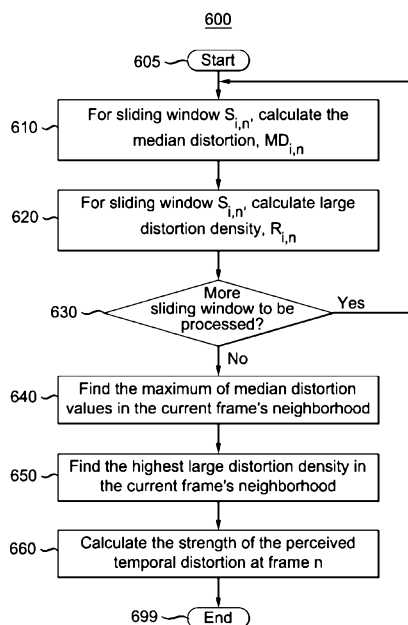
FIG. 6

(57) Abstract: Spatial distortion (i.e., when a frame is viewed independently of other frames in a video sequence) may be quite different from temporal distortion (i.e., when frames are viewed continuously). To estimate temporal distortion, a sliding window approach is used. Specifically, multiple sliding windows around a current frame are considered. Within each sliding window, a large distortion density is calculated and a sliding window with the highest large distortion density is selected. A distance between the current frame and the closest frame with large distortion in the selected window is calculated. Subsequently, the temporal distortion is estimated as a function of the highest large distortion ratio, the spatial distortion for the current frame, and the distance. In another embodiment, a median of spatial distortion values is calculated for each sliding window and the maximum of median spatial distortion values is used to estimate the temporal distortion.

TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

METHOD AND APPARATUS FOR CONTEXT-BASED VIDEO QUALITY ASSESSMENT

TECHNICAL FIELD

5   This invention relates to video quality measurement, and more particularly, to a method and apparatus for determining an objective video quality metric.

BACKGROUND

Different methods have been developed to estimate quality levels of individual frames in a video sequence. Given the estimated quality levels of individual frames
10   in the video sequence, the methods of obtaining an objective overall video quality metric that matches well with the perceived quality of the video sequence has been studied under the umbrella of temporal pooling strategies.

SUMMARY

The present principles provide a method for estimating visual quality of a
15   video sequence, comprising: accessing spatial distortion for frames in a plurality of sliding windows that include a current frame in the video sequence; determining a factor indicating at least one of a large distortion density and a representative artifact level for each sliding window responsive to respective spatial distortion for frames in the each sliding window; determining temporal distortion for the current frame
20   responsive to the determined factor; and determining the visual quality of the video sequence responsive to the temporal distortion of the current frame as described below. The present principles also provide an apparatus for performing these steps.

The present principles provide a method for estimating visual quality of a video sequence, comprising: accessing spatial distortion for frames in a plurality of sliding windows that include a current frame in the video sequence; determining a factor indicating at least one of a large distortion density and a representative artifact

5   level for each sliding window responsive to respective spatial distortion for frames in the each sliding window; determining a maximum of the factors for the plurality of sliding windows; determining temporal distortion for the current frame responsive to the maximum factor; and determining the visual quality of the video sequence responsive to the temporal distortion of the current frame as described below. The

10   present principles also provide an apparatus for performing these steps.

The present principles also provide a computer readable storage medium having stored thereon instructions for estimating visual quality of a video sequence according to the methods described above.

BRIEF DESCRIPTION OF THE DRAWINGS

15   FIG. 1A is a pictorial example depicting spatial artifact levels for individual frames in an exemplary video sequence, and FIG. 1B is a pictorial example depicting perceived temporal quality for individual frames in the exemplary video sequence.

FIGs. 2A, 2B, and 2C are pictorial examples depicting sliding windows used in video quality modeling, in accordance with an embodiment of the present principles.

20   FIG. 3A is a pictorial example depicting spatial artifact levels for individual frames in another exemplary video sequence, and FIGs. 3B and 3C are pictorial examples depicting dominant distortion in a frame's neighborhood and estimated

temporal distortion, respectively, in accordance with an embodiment of the present principles.

FIG. 4A is a pictorial example depicting spatial artifact levels for individual frames in another exemplary video sequence, and FIGs. 4B and 4C are pictorial examples depicting the highest large distortion density in a frame's neighborhood and estimated temporal distortion, respectively, in accordance with an embodiment of the present principles.

FIG. 5A is a pictorial example depicting spatial artifact levels for individual frames in another exemplary video sequence, and FIGs. 5B and 5C are pictorial examples depicting the highest large distortion density in a frame's neighborhood and estimated temporal distortion, respectively, in accordance with an embodiment of the present principles.

FIG. 6 is a flow diagram depicting an exemplary method for modeling temporal distortion at frame n, in accordance with an embodiment of the present principles.

FIG. 7 is a block diagram depicting an exemplary video quality monitor, in accordance with an embodiment of the present principles.

FIG. 8 is a block diagram depicting an exemplary video processing system that may be used with one or more implementations.

DETAILED DESCRIPTION

In the present application, we use "spatial artifact" to denote artifact perceived in a picture in a video sequence when the picture is viewed independently of other

pictures in the video sequence, and use "temporal artifact" to denote artifact that is perceived in a picture of a video sequence when pictures in the video sequence are continuously displayed. Similarly, we use "spatial distortion" or "spatial quality" to denote distortion or quality perceived in a picture when the picture is viewed

5   independently of other pictures in a video sequence, and use "temporal distortion" or "temporal quality" to denote distortion or quality that is perceived in a picture of a video sequence when pictures in the video sequence are continuously displayed.

When assessing spatial distortion, a picture is viewed independently of other pictures in the video sequence, for a period of time that is long enough for a viewer

10   to recognize image content and distortion. This is different from assessing temporal distortion, wherein pictures are continuously displayed.

In the present application, the term "frame" is used interchangeably with the term "picture." The terms "distortion" and "artifact" both refer to quality degradation in a frame, and are often used interchangeably.

15   Spatial distortion of frame n, denoted as ds(n), can be obtained by various image quality assessment methods, for example, but not limited to, a full-reference or no-reference method, and a method in a pixel domain or at a bitstream level. In one embodiment, spatial distortion can be estimated as the overall visible artifact level, caused by initial and/or propagated visible artifact, as disclosed in a commonly

20   owned PCT application, entitled "Video quality assessment at a bitstream level" by N. Liao, Z. Chen, and K. Xie (PCT/CN2012/078766, Attorney Docket No. PA120027), the teachings of which are specifically incorporated herein by reference.

Spatial artifact in pictures needs to last for a period of time so that eyes can fix

on and recognize it as artifact. When the pictures are part of a video sequence and each is displayed only for a very short period of time (for example, a period of 1/frame_rate when the video is played in real time), the perceived video distortion at the time instant of frame n, i.e., temporal distortion at frame n, dt(n), can be quite

5 different from spatial distortion of frame n, ds(n). This may be because temporal distortion dt(n) is affected by the content and distortion of its neighboring frames.

For an exemplary video sequence, FIG. 1A shows spatial artifact levels of individual frames in the video sequence, and FIG. 1B shows temporal quality of individual frames in the video sequence. More specifically, FIG. 1A shows spatial

10 artifact levels of the frames when the exemplary video sequence suffers from packet losses. The spatial artifact may be sporadic in the video sequence, for example, the spatial artifact seen at frames 74, 77, 215, 261, and 262. The spatial artifact may also occur in a burst, such as the artifact seen around frames 106-111.

FIG. 1B shows temporal quality when the frames of the video sequence are

15 displayed continuously, wherein score 100 corresponds to the best quality. Depending on the quality measurement method, the quality score may be at a different scale. The accurate curve of temporal quality may be obtained using a subjective test method, for example, but not limited to, Single Stimulus Continuous Quality Evaluation (SSCQE), as defined in ITU-R BT 500.

20 It can be seen from FIG. 1A that frames 74 and 77 have strong spatial artifact when these two frames are viewed independently. However, the artifact at these two frames becomes invisible when the video is displayed continuously, and thus, frames 74 and 77 are at the best quality level when viewed continuously as shown in FIG.

1B. Thus, strong spatial artifact may not always correspond to high temporal

distortion. On the other hand, one frame of a video sequence may appear to have

good quality when viewed independently of other frames, but may present very

strong temporal distortion (for example, motion jitter) when the video sequence is

5   displayed continuously. That is, small spatial distortion may not always correspond to

small perceived temporal distortion (i.e., higher temporal quality).

Thus, spatial distortion (i.e., when a picture is viewed independently of other

pictures in a video sequence) may be quite different from temporal distortion (i.e.,

when pictures are viewed continuously). The present principles provide a method

10   and apparatus for accurately modeling temporal quality from spatial distortion for

individual frames. Specifically, the present principles consider the context that affects

how a viewer identifies temporal distortion, wherein the context includes, for example,

but not limited to, the duration and the pattern of the distortion, and texture and

object's motion that are recognized by a viewer via watching the neighboring frames.

15   In one embodiment, given the spatial distortion, the perceived temporal

distortion can be modeled using a sliding window approach. As shown in FIGs. 2A-

2C, a sliding window of L_0 frames that includes frame n (denoted as $S_{i,n}$) starts at

frame $(n - i)$ and ends at frame $(n - i + L_0 - 1)$, $0 \leq i < L_0$. We may choose a

window length that approximates the period of time that is needed for a viewer's

20   eyes to recognize distortion. Accordingly, a sliding window includes $L_0 = \text{round}(T_0 \times$

$\text{frame\_rate})$ frames, where T_0 is a period of time that is needed for a viewer to

recognize distortion, and frame_rate is the frame rate of the video sequence. Artifact

existing outside the sliding windows is regarded as having little contribution to the

visibility of the temporal artifact of a current frame.

We observe that the perceived temporal distortion of a current frame is mainly affected by frames with large distortion (i.e., distortion level exceeds a certain threshold) that are close by. In the present principles, we provide different embodiments to estimate the impact of such clustered large spatial distortion in a

5   frame's neighborhood.

In one embodiment, we may use the maximum of median distortion values in the sliding windows to estimate the temporal distortion. Mathematically, for each sliding window $S_{i,n}$, we can calculate the median distortion of the frames in $S_{i,n}$ as:

$$MD_{i,n} = median\{ds(j), frame\ j \in S_{i,n}\}. \qquad (1)$$

10   A median function examines neighboring frames to decide whether or not the spatial distortion of the current frame is representative of its surroundings and rejects extreme distortion levels (outliers). That is, we may consider $MD_{i,n}$ as a representative distortion level for sliding window $S_{i,n}$. In other embodiment, we may use averaging or other filtering methods to obtain $MD_{i,n}$.

15   In all sliding windows that include frame n, the sliding window with larger distortion usually affects the visual quality more significantly. Thus, a maximum function can be used to identify the dominant distortion as the maximum of median distortion values among the sliding windows $\{S_{i,n}, 0 \leq i < L_0\}$:

$$m_n = max\{MD_{i,n}, \ 0 \leq i < L_0\}. \qquad (2)$$

20   Perceived temporal distortion for the current frame is observed to be largely determined by the dominant distortion in its neighborhood, rather than by the spatial distortion strength of the current frame. For example, when frame n has severe spatial distortion ds(n) which does not last long enough to be recognized, the

perceived temporal distortion is mainly affected by distortion in its neighborhood. Thus, a minimum function can be used to set the temporal distortion for the current frame. That is, we may set temporal distortion $dt(n)$ to the smaller one of the spatial distortion value and the maximum of median distortion values m_n for frame n as:

$$dt(n) = \min\{m_n, \ ds(n)\}. \tag{3}$$

FIG. 3A shows spatial artifact levels for frames in an exemplary video sequence, and FIGs. 3B and 3C show results after applying Eqs. (2) and (3), respectively. By using a median function, the maximum of median distortion values avoid extreme values as shown in FIG. 3B. Consequently, as can be seen from FIG. 3C, spatial artifact levels that are much larger than neighboring ones, for example, at frames 86, 125, and 166, are not present in the estimated temporal distortion. In addition, the estimated temporal distortion levels have smaller variations from frame to frame than the spatial distortion levels.

In another embodiment, to estimate temporal distortion, we may use the maximum value of the ratios of frames with large distortion (i.e., the distortion level exceeds a certain threshold) in all the sliding windows $\{S_{i,n}, 0 \leq i < L_0\}$. Since smaller distortion usually becomes invisible when the video sequence is played continuously, we only consider frames with large distortion. Mathematically, for each sliding window $S_{i,n}$, we can calculate the ratio between the number of frames with large distortion in $S_{i,n}$ and the total number of frames in $S_{i,n}$ as:

$$R_{i,n} = \frac{\Sigma_j \mu(ds(j))}{L_0}, \text{frame } j \in S_{i,n}, \tag{4}$$

where $\mu(x) = \begin{cases} 1, & x \geq U \\ 0, & \text{otherwise} \end{cases}$, and U is a spatial distortion threshold. For ease of notation, we denote the ratio between the number of frames with large spatial

distortion in a sliding window and the total number of frames in the sliding window as a large distortion density for the sliding window.

We observe that a sliding window with a higher density of larger distortion usually affects the visual quality more significantly. Thus, we calculate the highest ratio among all the sliding windows as:

$$w_n = \max\{R_{i,n}, 0 \leq i < L_0\}. \tag{5}$$

Then we can estimate the perceived temporal distortion at frame n as a weighted distortion:

$$dt(n) = w_n \times ds(n). \tag{6}$$

In addition, we observe that the distance between two nearest frames having large distortion also affects the perceived temporal distortion. That is, when two frames with large distortion are closer, the distortion becomes more visible to human eyes. Thus, the temporal distortion estimation may be improved by also considering the distance between frame n and a closest frame with large distortion. In one embodiment, Eq. (6) may be modified to become:

$$dt(n) = w_n \times ds(n)/f(dist(n)), \tag{7}$$

wherein dist(n) is the distance between frame n and the closest frame with large distortion in a sliding window corresponding to the highest large distortion density. If there is no other frame with large distortion in the corresponding sliding window, we set dist(n) to a very big value, for example, 1000. That is, when there is only one frame in the sliding window with large distortion, we consider the distortion as less visible and set dt(n) to a very small value. Different functions can be used to define f(dist(n)), for example, f(dist(n)) = dist(n), or f(dist(n)) = $(dist(n))^{\alpha}$, where α is a constant.

FIG. 4A shows spatial artifact levels for frames in an exemplary video

sequence, and FIGs. 4B and 4C show results after applying Eqs. (5) and (7),

respectively, wherein L_0 = 1second, U = (number of macro blocks per frame)/100.

The values of L_0 and U may vary with configurations, for example, with the GOP

5   length, video resolution, and the frame rate.FIG. 5A shows spatial artifact levels for

frames in another exemplary video sequence, and FIGs. 5B and 5C show results

after applying Eqs. (5) and (7), respectively.

From FIGs. 4B and 5B, we observe that when large distortion is more

clustered, the ratio becomes higher. That is, the ratio effectively reflects whether

10   and to which degree neighboring frames affect the perceived quality of the current

frame. Consequently, as can be seen from FIGs. 4C and 5C, the duration of frames

with clustered large distortion (for example, around frame 106 in FIG. 4A, around

frame 105 and frame 200 in FIG. 5A) affects the estimated temporal distortion levels.

That is, a shorter duration of distortion as shown in FIG. 4A results in estimated

15   temporal distortion that is much smaller than the spatial distortion, while a longer

duration of distortion as shown in FIG. 5A results in estimated temporal distortion

that is in the same order as the spatial distortion. This is consistent with the human

perception, where lasting distortion has more negative impact on the perceived

quality.

20   As discussed above, we consider contexts provided in neighboring frames to

estimate temporal distortion for frame n. In one embodiment, as shown in Eqs. (1)-

(3), we use the median distortion value. In another embodiment, as shown in Eqs.

(4)-(7), we use the large distortion density. In yet another embodiment, the median

distortion value and the large distortion density approaches can be combined to

estimate the temporal distortion for frame n as:

$$dt(n) = w_n \times \min\{m_n, \ ds(n)\}. \hspace{3cm} (8)$$

In the above, we estimate temporal distortion based on the human vision property that eyes need a period of time that is long enough to recognize artifact. The temporal distortion may also be affected by other factors, for example, but not limited to, motion jerkiness. Consequently, the temporal distortion estimated as above may need to be adjusted to consider other factors. One exemplary combination can be $dt'(n) = dt(n) + c \times dt_2(n)$, where $dt_2(n)$ is the distortion caused by motion jerkness.

Based on the temporal distortion for individual frames in a video sequence, $\{dt(n)\}$, the overall perceived distortion of the video sequence can be derived using a temporal pooling strategy. For example, a simple average operation can be used, that is, $\text{overall distortion} = \frac{1}{N}\sum_{n=1}^{N} dt(n)$.

FIG. 6 illustrates an exemplary method 600 for modeling temporal distortion at frame n, according to the present principles. Method 600 starts at step 605. For sliding window $S_{i,n}$, it calculates the median distortion value at step 610, for example, using Eq. (1), and calculates the large distortion density at step 620, for example, using Eq. (4). It checks whether more sliding window needs to be processed at step 630. If yes, it returns the control to step 610. Otherwise, at step 640, it calculates the maximum of median distortion values in all sliding windows for frame n, for example, using Eq. (2). At step 650, it calculates the highest large distortion density in all sliding windows for frame n, for example, using Eq. (5). At step 660, it estimates the temporal distortion for frame n, for example, using Eq. (3),(6), or (8).

The distance between frame n and the closest frame with large distortion may be considered at step 660, for example, using Eq. (7).Method 600 ends at step 699.

In method 600, both the maximum of median distortion values and the highest large distortion density are used to estimate temporal distortion. In another

5 embodiment, only the maximum of median distortion values is used to estimate the temporal distortion. That is, steps 620 and 650 are not needed, and step 660 estimates the temporal distortion based on the maximum of median distortion values, for example, using Eq. (3). In another embodiment, only the highest large distortion density is used to estimate the temporal distortion. That is, steps 610 and 640 are

10 not needed, and step 660 estimates the temporal distortion based on the highest large distortion density, for example, using Eq. (6) or Eq. (7).

Method 600 or its variations may proceed in a different order of steps, for example, step 620 may be performed before step 610, step 650 may be performed before step 640.

15 The video quality modeling methods according to the present principles can be applied to measure video quality when a video sequence suffers from compression artifacts and/or packet losses. The present methods may be used in conjunction with spatial artifacts obtained at a bitstream level or a pixel level. That is, the present methods can be applied at both a bitstream level and a pixel level.

20 The present principles may be used in a video quality monitor to measure video quality. FIG. 7 depicts a block diagram of an exemplary video quality monitor 700. The input of apparatus 700 may include a transport stream that contains the bitstream. The input may be in other formats that contains the bitstream.

Demultiplexer 710 obtains packet layer information from the bitstream.

Decoder 720 parses the input stream to obtain more information. Decoder 720 may

or may not reconstruct the pictures. In other embodiments, the decoder may

perform the functions of the demultiplexer.

5   Using the decoded information, the spatial artifact levels are estimated in

spatial artifact level estimator730. Based on the estimated parameters, temporal

distortion levels are estimated at temporal distortion estimator 740, for example,

using method 600. A quality predictor 750 then pools temporal distortion levels for

individual frames into a quality score for the video sequence. The quality predictor

10   750 may consider other types of artifacts and the property of human visual property.

The video quality monitor700 may be used, for example, in ITU-T

P.NBAMS(parametric non-intrusive bitstream assessment of video media streaming

quality) standard, which works on video quality assessment models in two

application scenarios, namely, IPTV and mobile video streaming, also called HR

15   (High Resolution)scenario and LR (Low Resolution)scenario respectively. The

difference between the two scenario ranges from the spatio-temporal resolution of

video content and coding configuration to transport protocols and viewing conditions.

The input to the P.NBAMS VQM (Video Quality Model) is coded video

bitstream with all transmission packet headers (UDP/IP/RTP or UDP/IP/RTP/TS).

20   The output is an objective MOS score (Mean Opinion Score).A major target

application of P.NBAMS work is to monitor video quality in a set-top box (STB) or

gateway. P.NBAMS mode 1 model only uses bitstream information, and mode 2

model may decode part or all of the video sequence, and the pixel information is

used for visual quality prediction in addition to parsing the bitstream information in order to improve the prediction accuracy.

Referring to FIG. 8, a video transmission system or apparatus 800 is shown, to which the features and principles described above may be applied. A processor 805 processes the video and the encoder 810 encodes the video. The bitstream generated from the encoder is transmitted to a decoder 830 through a distribution network 820. A video quality monitor may be used at different stages.

In one embodiment, a video quality monitor 840 may be used by a content creator. For example, the estimated video quality may be used by an encoder in deciding encoding parameters, such as mode decision or bit rate allocation. In another example, after the video is encoded, the content creator uses the video quality monitor to monitor the quality of encoded video. If the quality metric does not meet a pre-defined quality level, the content creator may choose to re-encode the video to improve the video quality. The content creator may also rank the encoded video based on the quality and charges the content accordingly.

In another embodiment, a video quality monitor850 may be used by a content distributor. A video quality monitor may be placed in the distribution network. The video quality monitor calculates the quality metrics and reports them to the content distributor. Based on the feedback from the video quality monitor, a content distributor may improve its service by adjusting bandwidth allocation and access control.

The content distributor may also send the feedback to the content creator to adjust encoding. Note that improving encoding quality at the encoder may not

necessarily improve the quality at the decoder side since a high quality encoded video usually requires more bandwidth and leaves less bandwidth for transmission protection. Thus, to reach an optimal quality at the decoder, a balance between the encoding bit rate and the bandwidth for channel protection should be considered.

5   In another embodiment, a video quality monitor 860 may be used by a user device. For example, when a user device searches videos in Internet, a search result may return many videos or many links to videos corresponding to the requested video content. The videos in the search results may have different quality levels. A video quality monitor can calculate quality metrics for these videos and

10   decide to select which video to store. In another example, the user device may have access to several error concealment techniques. A video quality monitor can calculate quality metrics for different error concealment techniques and automatically choose which concealment technique to use based on the calculated quality metrics.

        The implementations described herein may be implemented in, for example, a

15   method or a process, an apparatus, a software program, a data stream, or a signal. Even if only discussed in the context of a single form of implementation (for example, discussed only as a method), the implementation of features discussed may also be implemented in other forms (for example, an apparatus or program). An apparatus may be implemented in, for example, appropriate hardware, software, and firmware.

20   The methods may be implemented in, for example, an apparatus such as, for example, a processor, which refers to processing devices in general, including, for example, a computer, a microprocessor, an integrated circuit, or a programmable logic device. Processors also include communication devices, such as, for example,

computers, cell phones, portable/personal digital assistants ("PDAs"), and other devices that facilitate communication of information between end-users.

Reference to "one embodiment" or "an embodiment" or "one implementation" or "an implementation" of the present principles, as well as other variations thereof,

5   mean that a particular feature, structure, characteristic, and so forth described in connection with the embodiment is included in at least one embodiment of the present principles. Thus, the appearances of the phrase "in one embodiment" or "in an embodiment" or "in one implementation" or "in an implementation", as well any other variations, appearing in various places throughout the specification are not

10   necessarily all referring to the same embodiment.

Additionally, this application or its claims may refer to "determining" various pieces of information. Determining the information may include one or more of, for example, estimating the information, calculating the information, predicting the information, or retrieving the information from memory.

15   Further, this application or its claims may refer to "accessing" various pieces of information. Accessing the information may include one or more of, for example, receiving the information, retrieving the information (for example, from memory), storing the information, processing the information, transmitting the information, moving the information, copying the information, erasing the information, calculating

20   the information, determining the information, predicting the information, or estimating the information.

Additionally, this application or its claims may refer to "receiving" various pieces of information. Receiving is, as with "accessing", intended to be a broad term.

Receiving the information may include one or more of, for example, accessing the information, or retrieving the information (for example, from memory). Further, "receiving" is typically involved, in one way or another, during operations such as, for

5   example, storing the information, processing the information, transmitting the information, moving the information, copying the information, erasing the information, calculating the information, determining the information, predicting the information, or estimating the information.

As will be evident to one of skill in the art, implementations may produce a variety of signals formatted to carry information that may be, for example, stored or

10   transmitted. The information may include, for example, instructions for performing a method, or data produced by one of the described implementations. For example, a signal may be formatted to carry the bitstream of a described embodiment. Such a signal may be formatted, for example, as an electromagnetic wave (for example, using a radio frequency portion of spectrum) or as a baseband signal. The

15   formatting may include, for example, encoding a data stream and modulating a carrier with the encoded data stream. The information that the signal carries may be, for example, analog or digital information. The signal may be transmitted over a variety of different wired or wireless links, as is known. The signal may be stored on a processor-readable medium.

CLAIMS

1. A method for estimating visual quality of a video sequence, comprising:

accessing spatial distortion for frames in a plurality of sliding windows that include a current frame in the video sequence;

5   determining a factor indicating at least one of a large distortion density and a representative artifact level for each sliding window responsive to respective spatial distortion for frames in the each sliding window (610, 620);

determining temporal distortion for the current frame responsive to the determined factor (660); and

10   determining the visual quality of the video sequence responsive to the temporal distortion of the current frame.

2. The method of claim 1, further comprising:

determining a maximum of the factors for the plurality of sliding windows (640,

15   650), wherein the temporal distortion for the current frame is determined responsive to the maximum factor.

3. The method of claim 1, wherein the large distortion density is determined responsive to a ratio between the number of frames with large distortion in the each

20   sliding window and the total number of frames in the each sliding window.

4. The method of claim 1, wherein the temporal distortion for the current frame is determined as a product of spatial distortion for the current frame and the maximum factor.

25

5. The method of claim 1, wherein the representative artifact level is determined responsive to a median function of the spatial distortion for the each sliding window.
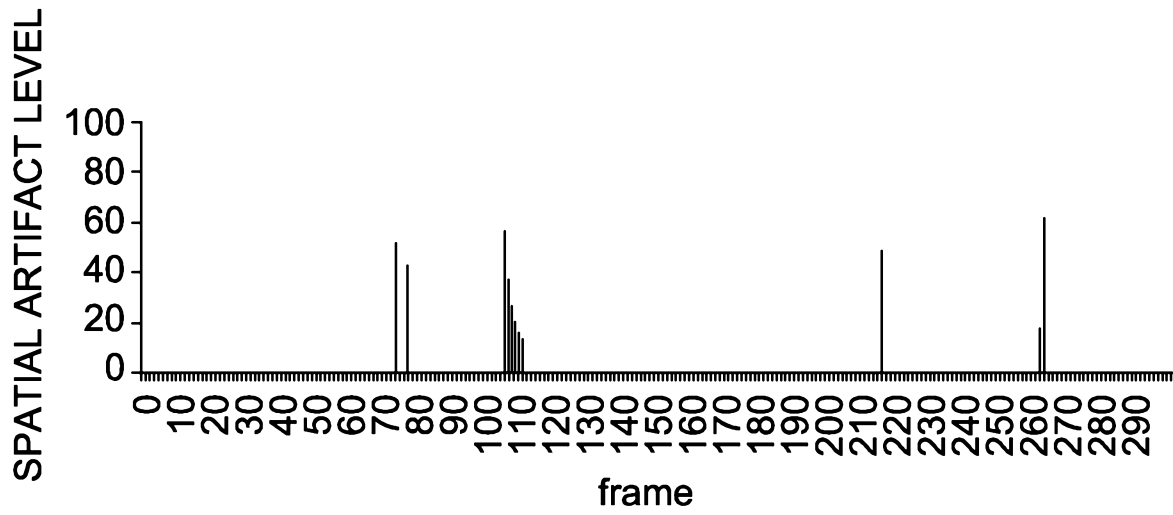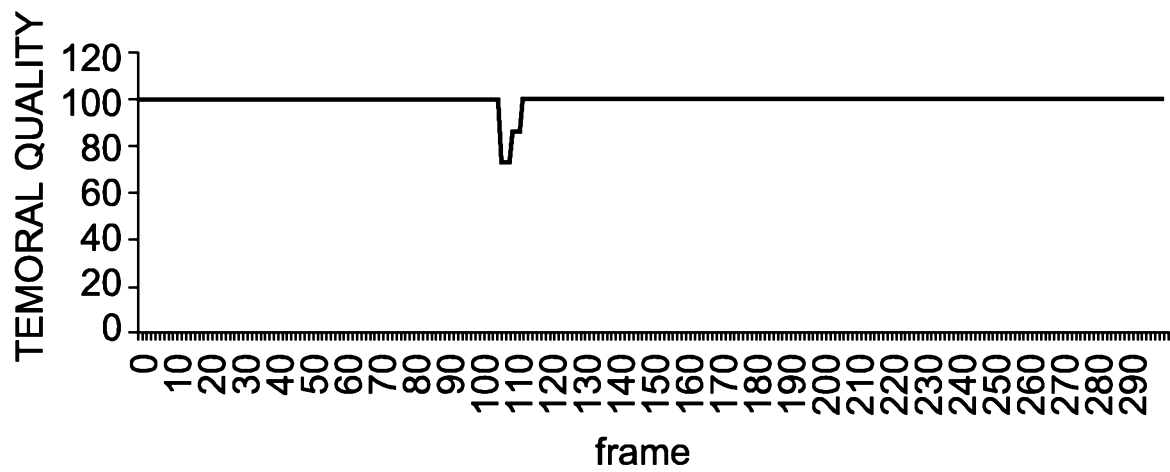
5   6. The method of claim 1, wherein the temporal distortion for the current frame is determined as a smaller one of spatial distortion for the current frame and the maximum factor.

7. The method of claim 1, further comprising:

10   performing at least one of monitoring quality of a bitstream, adjusting the bitstream in response to the determined visual quality, creating a new bitstream based on the determined visual quality, adjusting parameters of a distribution network used to transmit the bitstream, determining whether to keep the bitstream based on the determined visual quality, and choosing an error concealment mode at

15   a decoder.

8. An apparatus (700, 800) for estimating visual quality of a video sequence, comprising:

a spatial artifact level estimator (730) accessing spatial distortion for frames in

20   a plurality of sliding windows that include a current frame in the video sequence;

a temporal distortion estimator (740) determining a factor indicating at least one of a large distortion density and a representative artifact level for each sliding window responsive to respective spatial distortion for frames in the each sliding window and determining temporal distortion for the current frame responsive to the

25   determined factors for the plurality of sliding windows; and

a quality predictor (750) determining the visual quality of the video sequence responsive to the temporal distortion of the current frame.

9. The apparatus of claim 8, wherein the temporal distortion estimator (740)

5 determines a maximum of the factors for the plurality of sliding windows, wherein the temporal distortion for the current frame is determined responsive to the maximum factor.

10. The apparatus of claim 8, wherein the temporal distortion estimator (740)

10 determines the large distortion density responsive to a ratio between the number of frames with large distortion in the each sliding window and the total number of frames in the each sliding window.

11. The apparatus of claim 8, wherein the temporal distortion estimator (740)

15 determines the temporal distortion for the current frame as a product of the spatial distortion and the maximum factor.

12. The apparatus of claim 8, wherein the temporal distortion estimator (740) determines the representative artifact level responsive to a median function of the

20 spatial distortion for the each sliding window.

13. The apparatus of claim 8, wherein the temporal distortion estimator (740) determines the temporal distortion for the current frame as a smaller one of the spatial distortion and the maximum factor.

25

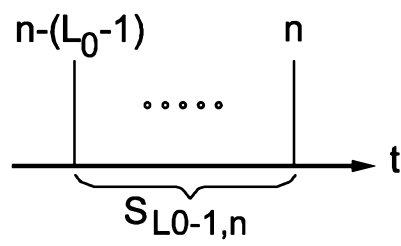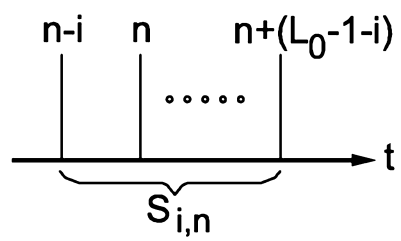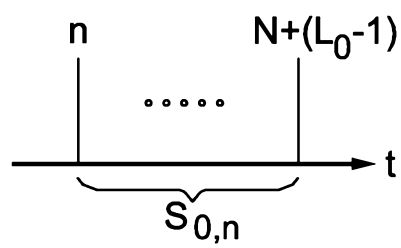14. The apparatus of claim 8, further comprising:

a video quality monitor (840, 850, 860) performing at least one of monitoring quality of a bitstream, adjusting the bitstream in response to the determined visual quality, creating a new bitstream based on the determined visual quality, adjusting

5   parameters of a distribution network used to transmit the bitstream, determining whether to keep the bitstream based on the determined visual quality, and choosing an error concealment mode at a decoder.
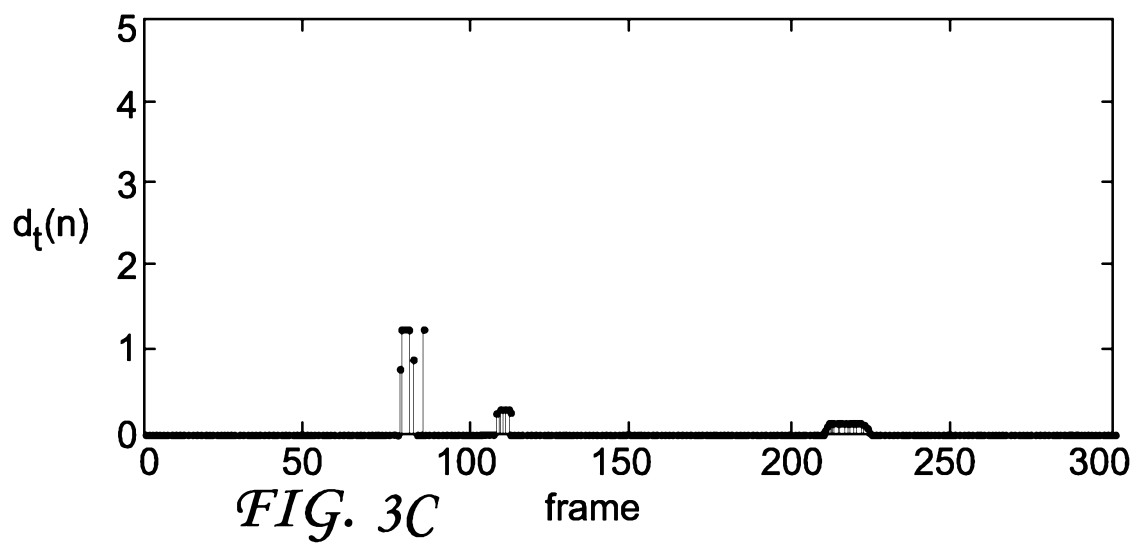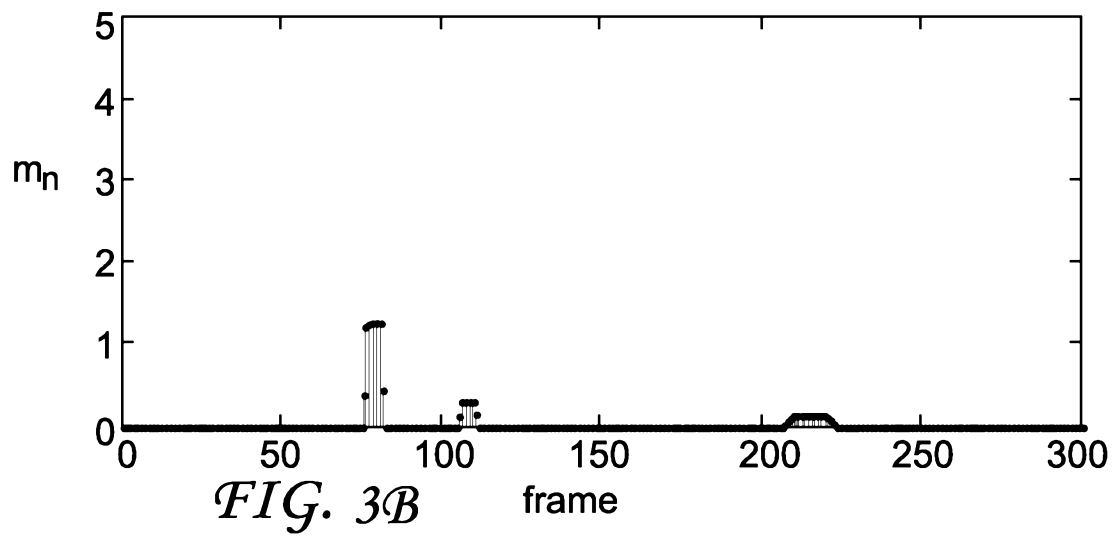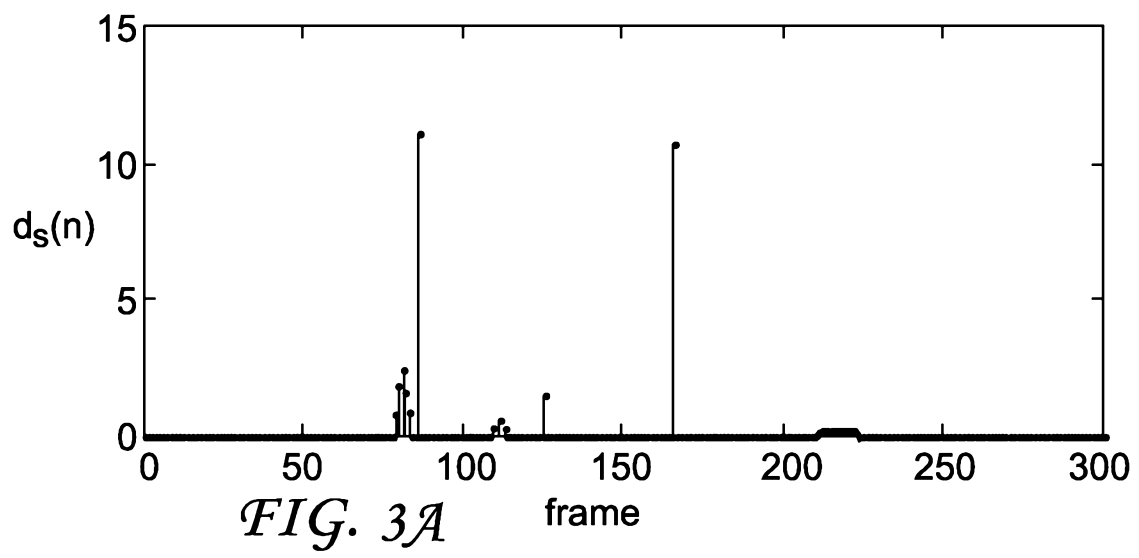
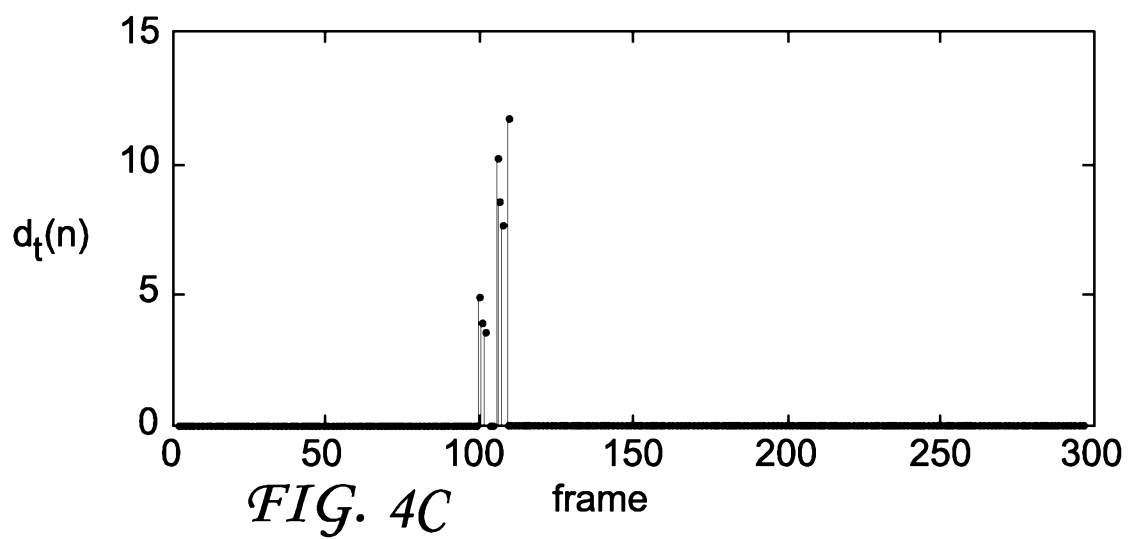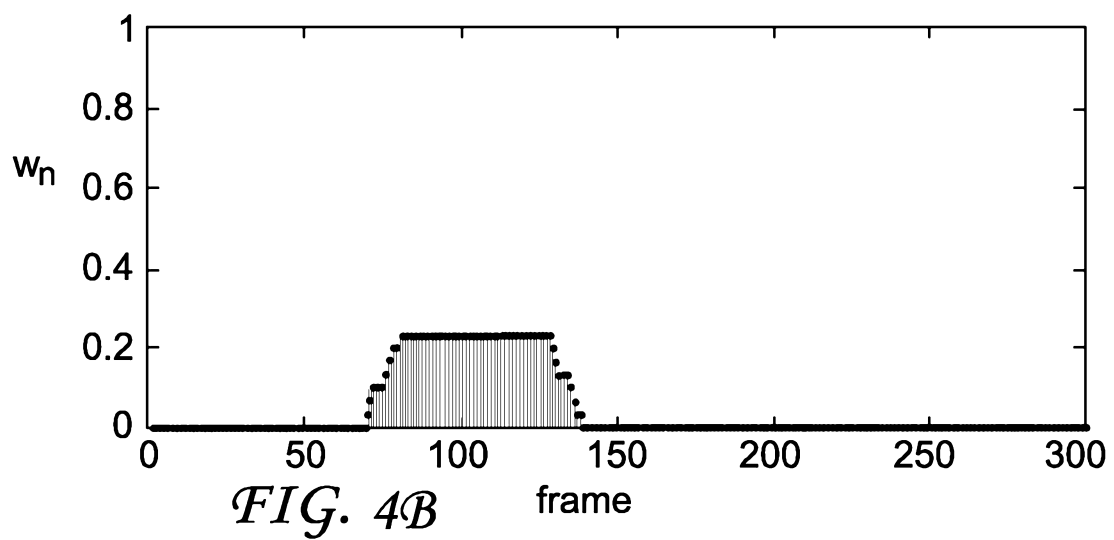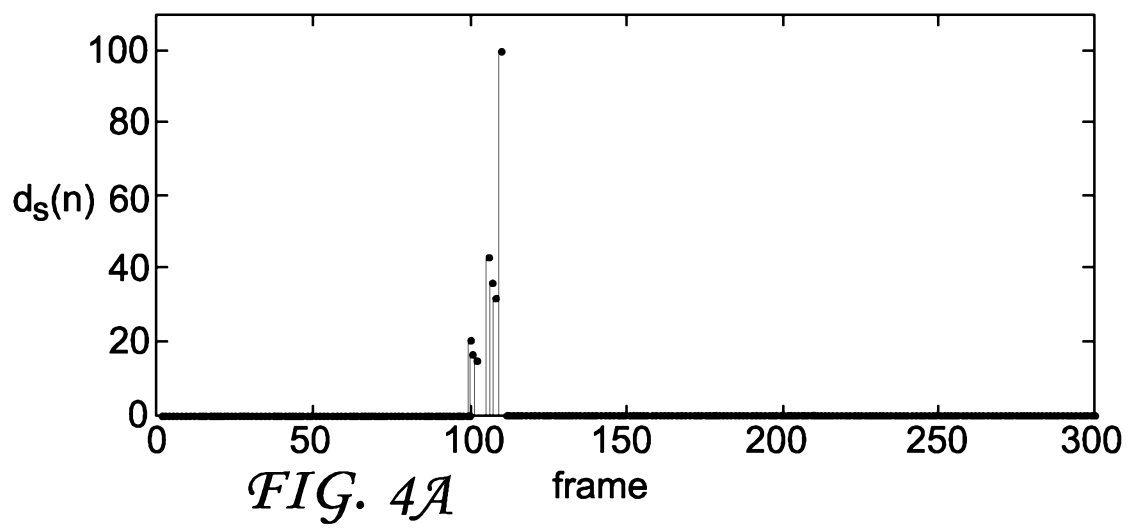15. A computer readable storage medium having stored thereon instructions

10   for estimating visual quality of a video sequence, according to claims 1-7.
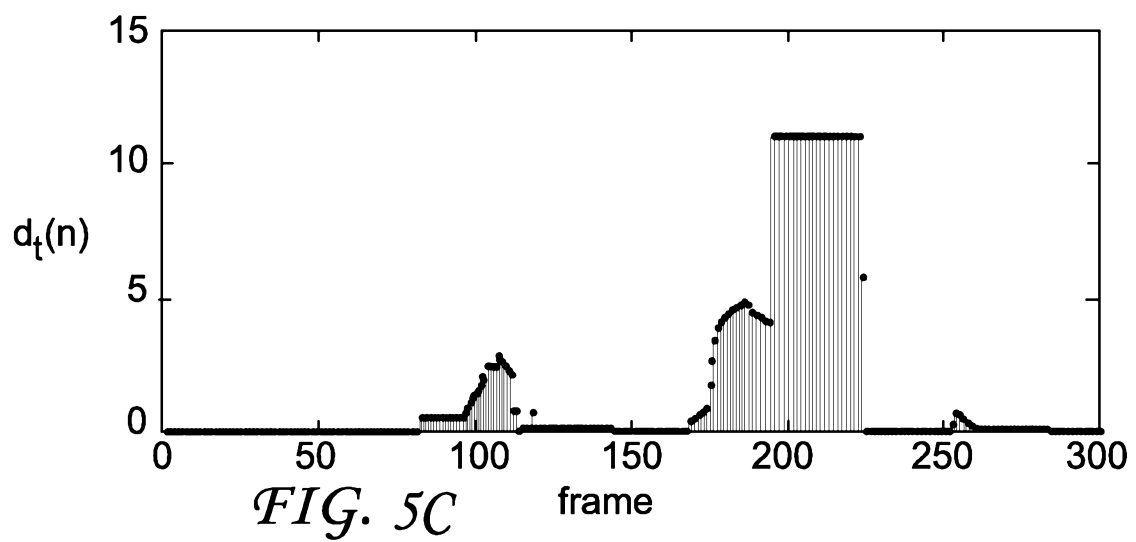
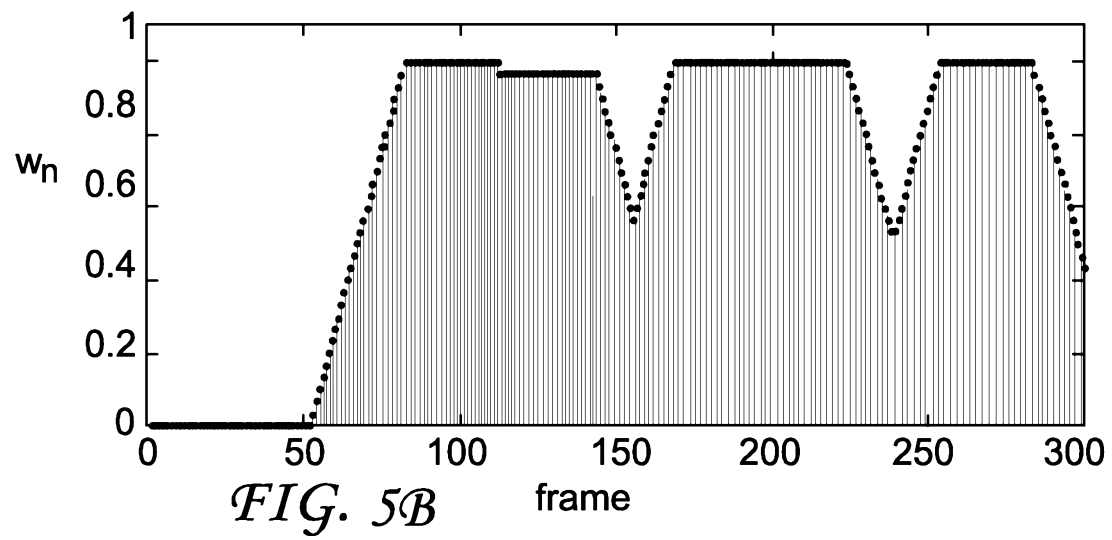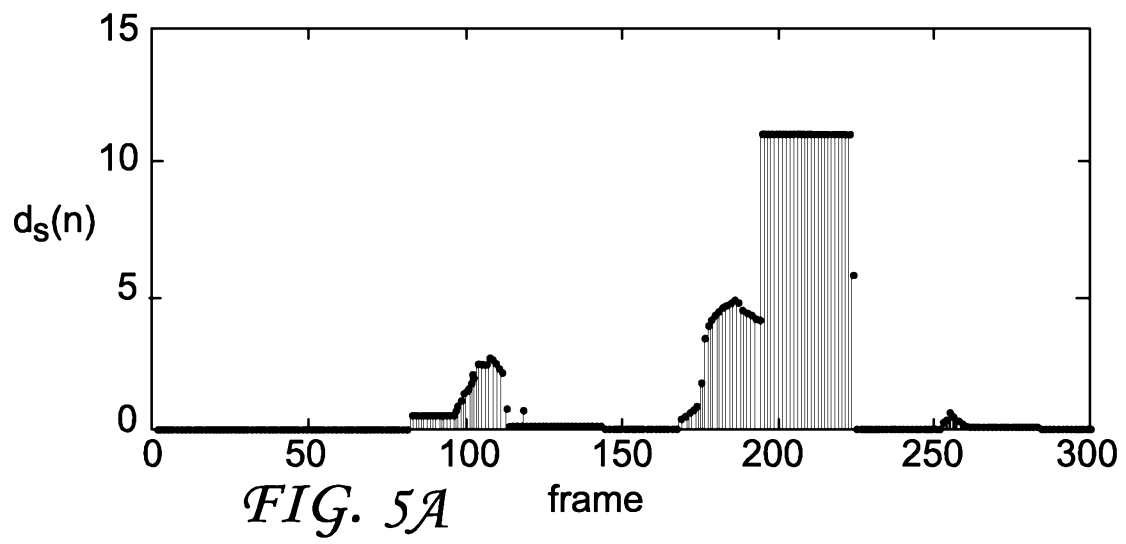FIG. 1A

FIG. 1B

$n-(L_0-1)$ n

$\circ \circ \circ \circ \circ$

t

$S_{L0-1,n}$

FIG. 2A

$n-i$ n $n+(L_0-1-i)$

$\circ \circ \circ \circ \circ$

t

$S_{i,n}$

FIG. 2B

n $N+(L_0-1)$

$\circ \circ \circ \circ \circ$

t

$S_{0,n}$

FIG. 2C

FIG. 3A

FIG. 3B

FIG. 3C

FIG. 4A

FIG. 4B

FIG. 4C

FIG. 5A

FIG. 5B

FIG. 5C

<u>600</u>

605 ⌇—(Start)

610 ⌇— For sliding window $S_{i,n}$, calculate the median distortion, $MD_{i,n}$

620 ⌇— For sliding window $S_{i,n}$, calculate large distortion density, $R_{i,n}$

630 ⌇— More sliding window to be processed? — Yes

No

640 ⌇— Find the maximum of median distortion values in the current frame's neighborhood

650 ⌇— Find the highest large distortion density in the current frame's neighborhood

660 ⌇— Calculate the strength of the perceived temporal distortion at frame n

699 ⌇—(End)

FIG. 6

FIG. 7

FIG. 8