



(51) International Patent Classification:

H04N 19/167 (2014.01)

(21) International Application Number:

PCT/FI2018/050213

(22) International Filing Date:

22 March 2018 (22.03.2018)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

1705063.4 30 March 2017 (30.03.2017) GB

(71) Applicant: **NOKIA TECHNOLOGIES OY** [FI/FI];
Karaportti 3, 02610 Espoo (FI).

(72) Inventor: **HOURUNRANTA, Ari**; Loutunkatu 29, 33560
Tampere (FI).

(74) Agent: **NOKIA TECHNOLOGIES OY** et al.; Ari Aarnio,
IPR Department, Karakaari 7, 02610 Espoo (FI).

(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, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, 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, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States (*unless otherwise indicated, for every kind of regional protection available*): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, 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, 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:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

(54) Title: VIDEO STREAMING

(57) Abstract: A method and system is disclosed for providing video data for streaming transmission to one or more remote devices. The video data may be arranged into video segments representing respective temporal portions of video content, having a given start time and duration. A plurality of versions of the same segment are provided, including a first version comprising video data having the given start time and duration, and at least one variant version in which the video data is divided into temporal sub-segments and in which one or more initial sub-segments contain substantially no video data and the other segment(s) contain video data for the remainder of the duration. The method may comprise receiving a request to transmit to a remote device a video segment for a given time which is referenced to the start time of the segment, and transmitting a selected one of the segment versions dependent on the time reference.



- 1 -
Video Streaming

Field of the Invention

This invention relates to video streaming, particularly view-dependent video streaming.

5

Background

It is known to provide data representing large scenes to users, only part of which can be seen at a given time. For example, in the field of virtual reality (VR), it is known to transmit video data representing a 360 degree video over a network to a user device, such as a VR headset.

10

The user will only see a portion of the 360 degree video, typically a 180 degree field-of-view, whilst the remainder of the video is in the background.

15

One of the main challenges for streaming such video is the high bandwidth requirements for acceptable subjective video quality. In the case of limited bandwidth, both seen and unseen portions of the video may have the same resolution. In view dependent streaming, only the portion that can be seen is streamed at a high quality, whilst the remaining, unseen portion is streamed at a lower quality.

20

One method for view-dependent delivery is to split the video images into tiles and download and decode in high quality only the tiles that are within the user's current field-of-view. The remaining part(s) of the video image are downloaded in very low quality, as a short-term fall back. Both the tiles and the background are streamed as segments that typically cover a duration of a few seconds. Segments typically start with a random access picture, for example an Intra, or I frame.

25

When switching to a new view in such content, it is important that latency is minimised or reduced. Latency may consist of two parts: the download time and tune-in time at the player. The tune-in time is due to the nature of video compression. If the time when the new target picture starts is not the same as the random access picture of the corresponding segment, the player will need to download additional pictures before the target picture. The download time may be optimised or improved by minimising the amount of data to download, and/or minimising the number of network operations required to download the data.

30

Summary of the Invention

35

A first aspect of the invention provides a method comprising: providing video data for streaming transmission to one or more remote devices, the video data being arranged into video segments representing respective temporal portions of video content, each video segment having a given start time and duration; providing one or more variant versions of

- 2 -

each segment in which the video data is divided into temporal sub-segments and in which one or more initial sub-segments contain substantially no video data and the other segment(s) contain video data from an offset time for the remainder of the duration; receiving a request to transmit to a remote device either the video segment or a variant version; and transmitting the requested video segment or variant version to the remote device.

The video data may be further arranged into groups of video segments, each group corresponding to a respective two-dimensional sub-area of a larger video image.

The request may be received responsive to a change in position of the remote device, or another device associated therewith, and the transmitting step comprises transmitting one or more new segments or variant versions corresponding to a new field-of-view.

The request may be associated with a switching time offset from the start of a current segment, and the transmitting step comprises transmitting the variant version which has a prior offset time closest to the switching time.

The method may further comprise, subsequent to transmitting one of the variant versions, transmitting the video segment corresponding to the next temporal portion in the event that no change in position is detected.

The video data may comprise a first variant version in which the first sub-segment contains substantially no video data, and one or more further variant versions in which an increasing number of successive sub-segment(s) contain substantially no video data.

The method may further comprise providing a data file comprising, for each variant version, the associated offset time and means for identifying each variant version to enable downloading thereof from a remote device, and transmitting the data file to one or more remote devices.

Each video segment may be divided into temporal sub-segments corresponding to those in each of the variant versions.

Each video segment may comprise a first reference frame substantially at the start of the segment, and the variant version(s) each comprise different reference frames substantially at the start of each sub-segment which contains video data.

The reference frame may be an intra frame.

Each video segment and its variant versions may further comprise one or more predicted frames based on the first reference frame, and wherein the variant versions further comprise the first reference frame to enable decoding of predicted frames in each sub-segment which contains video data.

5

The first reference frame may be provided substantially at the start of each variant version.

The video data may be virtual reality (VR) video data.

10

The video segments may be encoded using mono, stereo or multi-view coding.

The request may be received from a VR media player or VR headset.

The request may be a single HTTP request.

15

The video data may represent a 360 degree video image.

A second aspect of the invention provides a method comprising: receiving streaming video data from a remote device for displaying a current field-of-view, the streaming data being received as a plurality of video segments representing respective temporal portions of video content, each video segment having a given start time and duration; determining a new field-of-view; and transmitting a request to the remote device for receiving a variant version of one or more segments corresponding to the new field-of-view, the variant version being divided into temporal sub-segments and in which one or more initial sub-segments contain substantially no video data and the other segment(s) contain video data from an offset time for the remainder of the duration.

25

The video data may be further arranged into groups of video segments, each group corresponding to a respective two-dimensional sub-area of a larger video image.

30

The request may be transmitted responsive to a change in position of a user device, or another device associated therewith.

35

The method may further comprise receiving one or more new segments or variant versions corresponding to the new field-of-view.

The new field-of-view may be associated with a switching time which is offset from the start of a current segment, and the transmitting step comprises transmitting a request for the variant version which has a prior offset time closest to the switching time.

40

- 4 -

The method may further comprise, subsequent to requesting one of the variant versions, transmitting a further request for the video segment corresponding to the next temporal portion in the event that no change in position is detected.

- 5 The method may further comprise receiving a data file comprising, for each variant version of the segments, the associated offset time and means for identifying each variant version to enable requesting and receiving said variant version from the remote device.

10 Each received variant version(s) may comprise reference frames substantially at the start of each sub-segment which contains video data.

The reference frame may be an intra frame.

15 Each received variant version(s) may further comprise one or more predicted frames based on a first reference frame, and wherein the received variant version(s) further comprise the first reference frame to enable decoding of predicted frames in each sub-segment which contains video data.

20 The first reference frame may be received substantially at the start of each variant version. The video data may be virtual reality (VR) video data.

The method may be performed by VR media player or VR headset.

25 The transmitting step may comprise transmitting a single HTTP request.

The video data may represent a 360 degree video image.

30 A third aspect of the invention provides a computer program comprising instructions that when executed by a computer program control it to perform the method of providing video data for streaming transmission to one or more remote devices, the video data being arranged into video segments representing respective temporal portions of video content, each video segment having a given start time and duration; providing one or more variant versions of each segment in which the video data is divided into temporal sub-segments and
35 in which one or more initial sub-segments contain substantially no video data and the other segment(s) contain video data from an offset time for the remainder of the duration; receiving a request to transmit to a remote device either the video segment or a variant version; and transmitting the requested video segment or variant version to the remote device.

40

A fourth aspect of the invention provides a computer program comprising instructions that when executed by a computer program control it to perform the method of receiving streaming video data from a remote device for displaying a current field-of-view, the streaming data being received as a plurality of video segments representing respective temporal portions of video content, each video segment having a given start time and duration; determining a new field-of-view; and transmitting a request to the remote device for receiving a variant version of one or more segments corresponding to the new field-of-view, the variant version being divided into temporal sub-segments and in which one or more initial sub-segments contain substantially no video data and the other segment(s) contain video data from an offset time for the remainder of the duration.

A fifth aspect of the invention provides a non-transitory computer-readable storage medium having stored thereon computer-readable code, which, when executed by at least one processor, causes the at least one processor to perform a method, comprising:

providing video data for streaming transmission to one or more remote devices, the video data being arranged into video segments representing respective temporal portions of video content, each video segment having a given start time and duration; providing one or more variant versions of each segment in which the video data is divided into temporal sub-segments and in which one or more initial sub-segments contain substantially no video data and the other segment(s) contain video data from an offset time for the remainder of the duration; receiving a request to transmit to a remote device either the video segment or a variant version; and transmitting the requested video segment or variant version to the remote device.

A sixth aspect of the invention provides an apparatus, the apparatus having at least one processor and at least one memory having computer-readable code stored thereon which when executed controls the at least one processor: to provide video data for streaming transmission to one or more remote devices, the video data being arranged into video segments representing respective temporal portions of video content, each video segment having a given start time and duration; to providing one or more variant versions of each segment in which the video data is divided into temporal sub-segments and in which one or more initial sub-segments contain substantially no video data and the other segment(s) contain video data from an offset time for the remainder of the duration; to receiving a request to transmit to a remote device either the video segment or a variant version; and to transmit the requested video segment or variant version to the remote device.

- 6 -

A seventh aspect of the invention provides a non-transitory computer-readable storage medium having stored thereon computer-readable code, which, when executed by at least one processor, causes the at least one processor to perform a method, comprising: receiving streaming video data from a remote device for displaying a current field-of-view, the streaming data being received as a plurality of video segments representing respective temporal portions of video content, each video segment having a given start time and duration; determining a new field-of-view; and transmitting a request to the remote device for receiving a variant version of one or more segments corresponding to the new field-of-view, the variant version being divided into temporal sub-segments and in which one or more initial sub-segments contain substantially no video data and the other segment(s) contain video data from an offset time for the remainder of the duration.

An eighth aspect of the invention provides an apparatus, the apparatus having at least one processor and at least one memory having computer-readable code stored thereon which when executed controls the at least one processor: to receive streaming video data from a remote device for displaying a current field-of-view, the streaming data being received as a plurality of video segments representing respective temporal portions of video content, each video segment having a given start time and duration; to determining a new field-of-view; and to transmit a request to the remote device for receiving a variant version of one or more segments corresponding to the new field-of-view, the variant version being divided into temporal sub-segments and in which one or more initial sub-segments contain substantially no video data and the other segment(s) contain video data from an offset time for the remainder of the duration.

A ninth aspect of the invention provides an apparatus configured to perform the method of any preceding method definition.

Brief Description of the Drawings

The invention will now be described, by way of non-limiting example, with reference to the accompanying drawings, in which:

Figure 1 is a perspective view of a VR display system, useful for understanding the invention; Figure 2 is a block diagram of a computer network including the Figure 1 VR display system, according to embodiments of the invention; Figure 3a is a schematic top-plan view of a virtual space; Figure 3b is a schematic internal view of part of the Figure 3a virtual space; Figure 4 is a block diagram of components of a VR content provider forming part of the Figure 2 VR display system;

Figure 5 is a block diagram representing the arrangement of video segments groups and video segment versions;

Figure 6 is a schematic view of video segments and variants in accordance with embodiments of the invention;

5 Figure 7 is a schematic view of video segments and variants in accordance with further embodiments of the invention;

Figure 8 is a schematic view of video segments and variants in accordance with further embodiments of the invention;

10 Figure 9 is a flow diagram showing processing steps performed by the VR content provider in generating the segment variants, in accordance with embodiments;

Figure 10 is a block diagram of components of a VR media player forming part of the Figure 2 VR display system; and

Figure 11 is a flow diagram showing processing steps performed by the VR media player and the VR content player when consuming video content, in accordance with embodiments;

15

Detailed Description of Preferred Embodiments

Embodiments herein relate to video streaming, for example video streaming between a content provider and one or more user devices over a network, for example over an IP network such as the Internet.

20

More particularly, embodiments relate to view-dependent video streaming, where the data that is streamed from the source of the video content to a user end system is dependent on the position or orientation of the user.

25 In some embodiments, the video stream may represent a part of an overall, wide-angle video scene. For example, the video scene may cover a field which is greater than a viewer's typical field-of-view, e.g. greater than 180°. Therefore, embodiments are particularly suited to applications where a user may consume and/or interact with an overall video scene greater than 180° and possibly up to 360°.

30

One use case is virtual reality (VR) content whereby video content is streamed to a VR display system. As is known, the VR display system may be provided with a live or stored feed from a video content source, the feed representing a virtual reality space for immersive output through the display system. In some embodiments, audio is provided, which may be spatial
35 audio.

Nokia's OZO (RTM) VR camera is an example of a VR capture device which comprises a camera and microphone array to provide VR video and a spatial audio signal, but it will be

appreciated that the embodiments are not limited to VR applications nor the use of microphone arrays at the video capture point.

Figure 1 is a schematic illustration of a VR display system 1. The VR system 1 includes a VR headset 20, for displaying visual data in a virtual reality space, and a VR media player 10 for rendering visual data on the VR headset 20.

In the context of this specification, a virtual space is any computer-generated version of a space, for example a captured real world space, in which a user can be immersed. The VR headset 20 may be of any suitable type. The VR headset 20 may be configured to provide VR video and audio content to a user. As such, the user may be immersed in virtual space.

The VR headset 20 receives visual content from a VR media player 10. The VR media player 10 may be part of a separate device which is connected to the VR headset 20 by a wired or wireless connection. For example, the VR media player 10 may include a games console, or a PC configured to communicate visual data to the VR headset 20.

Alternatively, the VR media player 10 may form part of the display for the VR headset 20.

Here, the media player 10 may comprise a mobile phone, smartphone or tablet computer configured to play content through its display. For example, the device may be a touchscreen device having a large display over a major surface of the device, through which video content can be displayed. The device may be inserted into a holder of a VR headset 20. With these headsets, a smart phone or tablet computer may display visual data which is provided to a user's eyes via respective lenses in the VR headset 20. The VR display system 1 may also include hardware configured to convert the device to operate as part of VR display system 1. Alternatively, VR media player 10 may be integrated into the VR display device 20. VR media player 10 may be implemented in software. In some embodiments, a device comprising VR media player software is referred to as the VR media player 10.

The VR display system 1 may include means for determining the spatial position and/or orientation of the user's head. Over successive time frames, a measure of movement may therefore be calculated and stored. Such means may comprise part of the VR media player 10. Alternatively, the means may comprise part of the VR display device 20. For example, the VR display device 20 may incorporate motion tracking sensors which may include one or more of gyroscopes, accelerometers and structured light systems. These sensors generate position data from which a current visual field-of-view (FOV) is determined and updated as the user changes position and/or orientation. The VR display device 20 will typically

comprise two digital screens for displaying stereoscopic video images of the virtual world in front of respective eyes of the user, and also two speakers for delivering audio, if provided from the VR system. The embodiments herein, which primarily relate to the delivery of VR content, are not limited to a particular type of VR display device 20.

5

The VR display system 1 may be configured to display visual data to the user based on the spatial position of the display device 20 and/or the orientation of the user's head. A detected change in spatial position and/or orientation, i.e. a form of movement, may result in a corresponding change in the visual data to reflect a position or orientation transformation of the user with reference to the virtual space into which the visual data is projected. This allows VR content to be consumed with the user experiencing a 3D VR environment.

10

15

The VR display device 20 may display non-VR video content captured with two-dimensional video or image devices, such as a smartphone or a camcorder, for example. Such non-VR content may include a framed video or a still image. The non-VR source content may be 2D, stereoscopic or 3D. The non-VR source content includes visual source content, and may optionally include audio source content. Such audio source content may be spatial audio source content. Spatial audio may refer to directional rendering of audio in the virtual space such that a detected change in the orientation of the user's head may result in a corresponding change in the spatial audio rendering to reflect an orientation transformation of the user with reference to the virtual space in which the spatial audio data is rendered. The display of the VR display device 20 is described in more detail below.

20

25

The angular extent of the virtual environment observable through the VR display device 20 is called the visual field of view (FOV) of the display device 20. The actual FOV observed by a user depends on the inter-pupillary distance and on the distance between the lenses of the headset and the user's eyes, but the FOV can be considered to be approximately the same for all users of a given display device when the display device is being worn by the user.

30

Figure 2 shows a typical VR system 1, comprising the above-described media player 10 and VR display device 20. A remote content provider 30 may store and transmit streaming video data which, in the context of embodiments, is VR video for display to the VR display device 20. Responsive to receive or download requests sent by the media player 10, the content provider 30 streams the VR data over a data network 40, which may be any network, for example an IP network such as the Internet. Streaming may be by means of the MPEG-DASH standard but it not limited to such.

35

The remote content provider 30 may or may not be the location or system where the VR video is captured and processed.

5 The video content for the overall video scene, for example a 360 degree video scene, may be arranged as a series of two-dimensional areas or tiles, each representing a respective spatial part of the scene. Therefore, the VR media player 10 may only download from the content provider 30 those tiles which are within the current FOV, at least at a high quality.

10 Each tile may be represented by a video segment, which is a temporal portion of the video content, typically in the order of seconds, e.g. two seconds or similar. Therefore, for a given tile, multiple segments may be downloaded, buffered and then decoded and rendered in a sequential order.

15 Therefore, in some embodiments, there may be provided a plurality of groups of video segments, each group corresponding to a respective sub-area of the larger video content, and each segment within a group representing different temporal portions of the video content. The video content is therefore divided in both the spatial and temporal domains.

20 Figure 3a is a schematic top plan view representing a 360° view field 50 in relation to a user 55 wearing the VR display device 20. Based on the user's position and the orientation of the VR display device 20, only a current FOV 60 may be streamed to the media player 10, i.e. the portions or segments of the video scene between the bounding lines 57. Position signals from the VR display device 20 are transmitted to the media player 10 which determines the FOV 60 and requests downloading of the segments corresponding to the FOV 60. When the
25 segments corresponding to the FOV 60 are received, the media player 10 decodes and renders said segments to the VR display device 20.

30 Figure 3b shows, for example, a plurality of segments 80a – 80h when rendered to the VR display device 20 from the user's perspective. These may be termed “first segments” in that they represent the current FOV 60. Each segment 80a – 80h is effectively a tile representing a respective two-dimensional region of the FOV 60. Each segment 80a - 80h may represent video data lasting several seconds in length.

35 For the avoidance of doubt, the term segment used herein refers to video data representing a sub-portion of an overall image for a time interval.

Embodiments herein relate to pull-mode methods, i.e. where the media player 10 informs the content provider 30 of the video data it wants to download next, e.g. based on its position, or

predicted future position or FOV. MPEG-DASH is an example of such a pull-mode technology.

In MPEG-DASH, there is provided a manifest file, or Media Presentation Description (MPD), which is an XML file that provides information about the available media streams. The manifest file can be used to split the video stream in the spatial domain into adaptation sets and representations. Adaptation sets enable the grouping of different multimedia components that logically belong together. For example, components with the same codec, language, resolution etc. could be within the same adaptation set. This enables the client, e.g. the media player 10, to eliminate a range of media components that do not fulfil its requirements.

Representations define interchangeable versions of the respective content, e.g. different resolutions, bitrates etc. Although one single representation may provide enough information to provide a playable stream, multiple representations may give the client the possibility of adapting the media stream to current network conditions and bandwidth requirements.

The representations may provide definitions of time-wise attributes for the video, e.g. the start time of segments, their duration and random access periods. Regarding the latter, a random access period is the time when a random access (or reference) image occurs within the segment. A reference image is typically an intra frame (or I frame) from which later frames can be de-compressed.

In the case of view-dependent delivery (VDD), there may be plural adaptation sets presented in the manifest, each representing a sub-portion of the 360 degree view. In other words, each adaptation set corresponds to a group of segments. Each adaptation set may, e.g. provide different bit-rate variants of the video, but representations within an adaptation set or group will represent the same video content.

A combination of representations, one per adaptation set, may be downloaded and decoded to form a view for the user.

The manifest is however a template. Hence, it cannot contain references to byte indices since, due to the nature of video compression, they vary within the representation. Instead, it can give references in time, based on given timescale attributes.

In MPEG-DASH systems and alike, each video segment is fetched with a network request (e.g. HTTP in DASH). Typically, in low latency systems the duration of a segment is two seconds, or similar. The shorter the segments are, the more there is overhead, caused by I frames, and the processing of HTTP requests.

5

Embodiments herein provide methods for generating, providing and receiving video segments in streaming applications, for example VR streaming, and the methods are particularly suited to situations when view orientation changes and the media player 10 needs to download and tune-in to a new video stream. The motivation is to reduce latency when switching views but also avoiding additional overhead.

10

In overview, the methods provide, for each segment, one or more variants as separate downloadable streams, each downloadable from the content provider 30 by the VR media player 10 using a single fetch or download request, e.g. an HTTP request. Hence, the VR media player 10 may quickly select the best variant for a given time instant, instead of issuing multiple fetch requests.

15

This is in contrast with a known technique whereby the VR media player 10 must first fetch metadata from the start of a segment in order to read the segment's index table, before reading the byte ranges of sub-segments and fetching one or more sub-segments with a byte-range HTTP request. This would involve a round-trip time of two HTTP requests. In some cases, more than two HTTP requests are needed.

20

In some embodiments, each segment comprises multiple random access points, e.g. reference or I frames.

25

Figure 4 is a schematic diagram of components of the content provider 30, or a computer system associated with the content provider. The content provider 30 may have a controller 100, RAM 102, a memory 104, and, optionally, hardware keys 106 and a display 108. The content provider 30 may comprise a network interface 110 for connection to the network 40, e.g. a modem which may be wired or wireless. The network interface 110 may therefore be used to receive download requests from the VR display system 1 and to stream data to the VR display system 1. A segment database 116 is also provided, for storing video data for streaming transmission to external devices, such as the VR display system 1.

30

The controller 100 is connected to each of the other components in order to control operation thereof.

35

The memory 104 may be a non-volatile memory such as read only memory (ROM), a hard disk drive (HDD) or a solid state drive (SSD). The memory 104 stores, amongst other things, an operating system 112 and may store software applications 114. The RAM 102 is used by the controller 100 for the temporary storage of data. The operating system 112 may contain code which, when executed by the controller 100 in conjunction with the RAM 102, controls operation of each of the hardware components of the content provider 30.

The controller 100 may take any suitable form. For instance, it may be a microcontroller, plural microcontrollers, a processor, or plural processors.

The content provider 30 may be a standalone computer, a server, a console, or a network thereof. The content provider 30 may communicate with the VR display system 1 in accordance with one or more software applications 112 in accordance with steps to be described later on.

In some embodiments, the content provider 30 may also be associated with external software applications. These may be applications stored on a remote server device and may run partly or exclusively on the remote server device. These applications may be termed cloud-hosted applications. The content provider 30 may be in communication with the remote server device in order to utilize the software application stored there.

Figure 5 is a schematic view of the segment database 116 within which are stored groups of segments 120 (corresponding to the different spatial tiles or areas). Within each group 120 are provided the individual segments 122 each of which relates to a different temporal part of the video stream, each segment having a duration of approximately 2 seconds. Each segment 122 has an associated start time and duration. Additionally, each segment 122 is associated with a plurality of versions 124, as will be described with reference to Figure 6. A manifest file 126 is also provided, the purpose of which will be described later on.

Figure 6 shows the different versions 124 of a given video segment 122. Each version 124 may be generated at the content provider 30, or received as separate files from an external source. The versions comprise a first version 130, and three variant versions (hereafter “variants”) 132, 134, 136.

The first version 130 comprises the video data for the entire segment duration, e.g. 2 seconds, and is effectively the original video segment divided into sub-segments, which may be of equal duration. In the shown example, there are four sub-segments 141 - 144, each having video data of duration 0.5 seconds.

The variants 132, 134, 136 differ from the first version 130 in that one or more of the initial sub-segments 141 – 143 do not contain or represent video data, i.e. they are blank or skipped. For example, the first variant 132 contains no video data corresponding to the first sub-segment 141. The second variant 134 contains no video data corresponding to the first and second sub-segments 141, 142. The third variant 136 contains no video data corresponding to the first to third sub-segments 141, 142, 143.

Each of the first version 130 and the three variants 132, 134, 136 begin with a set of metadata 138, respectively identifying each variant. The metadata 138 further contains information about the sub-segments in the first version and the variants 132, 134, 136, for example the location of the sub-segments as byte offsets. For the variants 132, 134, 136 the metadata may only contain information about the included sub-segments.

The manifest file 126 may provide a template mode file, comprising the start time for the overall streaming representation and the segment duration, which may be constant for all segments of a group of video segments. The manifest file 126 may also comprise a template for the URL, comprising a common part and a segment index which will make the URL unique. For example, the URL may be of the form {baseurl}/segment.tile1~\$index\$.mp4 where {baseurl} gives the server's IP address and path, and is common to all streams of the particular streaming representation. The index part is replaced with the segment index that the VR media player 10 wishes to download, which in embodiments herein may also include an additional field for variants. In use, the VR media player 10 may fetch the appropriate segment using the time difference between the current time and the start time, and the segment durations to derive the segment index. To enable the media player 10 to select the appropriate variant, if needed, a property may be provided in the manifest file 126 that indicates how to map a time offset to the corresponding variant URL. For example, there may be provided a property that lists the offsets and, for each, a paired ID. The ID is that which is appended to the template URL.

Further, in all versions, each sub-segment 141 – 144 that contains video data starts with a random access picture 150, for example an intra frame.

Each of the versions 130, 132, 134, 136 may be downloaded by the VR media player 10 of the VR display system 1 using a single request, for example a single HTTP request in the form indicated above.

In embodiments, the VR media player 10 may download a selected one of the segment versions 130, 132, 134, 136 using a single request based on where in the media timescale the download is needed. This is particularly useful in the context of view-dependent streaming.

- 15 -

For example, initially, the VR media player 10 may be downloading the same segment stream because the VR headset 20 may not be moving, or is moving very little. Therefore, the VR media player 10 may request the first version 130 because it contains video data in all of the sub-segments 141 – 144, and so it downloads and buffers the segments, and then decodes and renders them in the sequential decoding order. However, if the VR headset 20 moves to a new FOV, then it will have to switch to downloading new segments for the new FOV, i.e. from a different group of segments.

Depending on the time when this occurs, one of the variants 132, 134, 136 may be downloaded instead. For example, if the switch occurs at a time instant corresponding to 1.2 seconds from the start time of the segment, i.e. $T_1 + 1.2$ seconds, the VR media player 10 will need video data corresponding to the second and third sub-segments 143, 144 because the first and second sub-segments 141, 142 have already passed in time. Hence, the VR media player 10 will request from the content provider 30 the second variant 134 for downloading, decoding and rendering. It may be generalised therefore that for a given offset time the needed variant version is that which has a prior offset time closest to the switching time.

If the VR headset 20 does not move subsequently, then the VR media player 10 may select (for the next segment in time) the first version 130 because video data is contained in all sub-segments 141 – 144.

Referring to Figure 7, a further embodiment is shown whereby the content provider 30 may generate or receive different versions 160 of a given video segment, in the situation where the video stream uses so-called Dependent Random Access Point (DRAP) pictures. The versions comprise a first version 170, and three variants 172, 174, 176.

In overview, DRAP schemes use special inter-predicted frames within the segment, which are predicted from the first Intra (I) frame of the segment. The aim is to increase compression efficiency.

Similar to the Figure 6 example, the first version 170 comprises the video data for the entire segment duration, e.g. 2 seconds, and is effectively divided into sub-segments, which may be of equal duration. In the shown example, there are four sub-segments 181 - 184, each having video data of duration 0.5 seconds.

The variants 172, 174, 176 differ from the first version 170 in that one or more of the initial sub-segments 181 – 183 do not contain or represent video data, i.e. they are blank or skipped. For example, the first variant 172 contains no video data corresponding to the first sub-

segment 181. The second variant 174 contains no video data corresponding to the first and second sub-segments 181, 182. The third variant 176 contains no video data corresponding to the first to third sub-segments 181, 182, 183. Each of the first version 170 and the three variants 172, 174, 176 begin with a set of metadata 138, respectively identifying each variant.

5

Different from the Figure 6 embodiment, if the video data contains DRAP, inter-predicted frames, then the VR media player 10 will also require the initial Intra frame which occurs in, or is associated with, the first sub-segment 181. Accordingly, the provided variants 172, 174, 176 include an additional sub-segment 192 at the beginning, which includes only the I frame of the first sub-segment 181, indicated I₀, and no other video content from the sub-segments preceding the starting sub-segment of the variant. This enables the VR media player 10 to decode the video content.

10

15

Referring to Figure 8, a further embodiment is shown whereby the content provider 30 may generate or receive different versions 200 of a given video segment, in the situation where the content provider has a relatively large amount of encoding power and quality is to be optimised.

20

The versions comprise a first version 202, and three variants 204, 206, 208.

Similar to the Figure 6 example, the first version 202 comprises the video data for the entire segment duration, e.g. 2 seconds, but is not divided into sub-segments with additional random access points. The variants 172, 174, 176 may then have additional random access points to enable a faster tune-in time.

25

In all the above, a greater or fewer number of variants may be provided, and a greater or fewer number of sub-segments. Although Figures 6 to 8 represent a single segment, it will be appreciated that in practice there will be a large number of segments representing video content both spatially and temporally.

30

Figure 9 is a flow diagram indicating processing steps performed by the content provider 30 under control of the software application 114 for generating the segment versions.

35

In a first step 9.1, the video segments for each spatial area are received. In a second step 9.2, each video segment is taken in turn and, in step 9.3, a first version of the segment is created, for example by dividing the segment into sub-segments as shown in Figures 6 and 7. A next step 9.4 comprises creating, or generating, one or more variant segment versions, as shown in Figures 6, 7 and 8. A next step 9.5 comprises storing the segment versions in the segment

database 116. A final step 9.6 comprises creating or updating a manifest file, e.g. the XML file for informing the VR media player 10 about the variants and which offset times each variant corresponds to.

- 5 It will be appreciated that certain steps of the Figure 9 method may be reordered and/or performed in parallel.

For completeness, Figure 10 is a schematic diagram of components of the media player 10. The media player 10 may have a controller 230, RAM 232, a memory 234, and, optionally,
10 hardware keys 236 and a display 238. The media player 10 may comprise a network interface 240 for connection to the network 40, e.g. a modem which may be wired or wireless. The media player 10 also comprises a wired or wireless port for transmitting and receiving signals with the VR display device 10. The input signals from the display device 10 will be position signals, indicative of user position/orientation and from which can be computed
15 instantaneous and/or averaged movement over time. The output signals to the display device 10 will be the decoded and rendered video segment data. The controller 230 is connected to each of the other components in order to control operation thereof.

The memory 234 may be a non-volatile memory such as read only memory (ROM), a hard
20 disk drive (HDD) or a solid state drive (SSD). The memory 234 stores, amongst other things, an operating system 242 and may store software applications 244. The RAM 232 is used by the controller 230 for the temporary storage of data. The operating system 242 may contain code which, when executed by the controller 230 in conjunction with the RAM 232, controls operation of each of the hardware components of the media player 10.

25 The controller 230 may take any suitable form. For instance, it may be a microcontroller, plural microcontrollers, a processor, or plural processors.

The media player 10 may be a standalone computer, a server, a console, or a network thereof.
30 In some embodiments, the media player 10 may be provided as part of the VR display device 20. In such cases, both the media player 10 and VR display device 20 may be collectively referred to as a user device. The media player 10 may communicate with the content provider 30 and the VR display device 20 in accordance with one or more software applications 242.

35 In some embodiments, the media player 10 may also be associated with external software applications not stored on the media player. These may be applications stored on a remote server device and may run partly or exclusively on the remote server device. These

applications may be termed cloud-hosted applications. The media player 10 may be in communication with the remote server device in order to utilize the software application stored there.

- 5 Figure 11 is a flow diagram indicating processing steps performed by the media player 10 under control of the software application 242 for requesting the segment versions for segments of a VR video stream. Certain steps which are performed by the content provider 30 are also shown for ease of explanation, although it will be appreciated that these steps are not performed by the media player 10.

10

In a first step 11.1, the media player 10 receives a manifest file from the content provider 30. As mentioned previously, the manifest file may be an XML file indicating each segment and the associated variants, including offset times “x” for the variants. In a second step 11.2, the media player 10 identifies the current FOV for the VR headset 20 (or other user device). In a
15 third step 11.3, the media player 10 requests segments corresponding to the current FOV. For ease of illustration, we assume that no movement has taken place immediately prior to this step, and hence the first version of each required segment is requested. The request may be by means of a single HTTP request.

- 20 At the content provider 30, responsive to the request, the first version segments are transmitted to the media player 10 in step 11.4.

In step 11.5, the media player 10 receives the first version segments for buffering, decoding and rendering to the VR headset 20.

25

In step 11.6, which assumes that some movement of the VR headset 20 has occurred, a new FOV is determined for a time point (or switch point, given that a switch between segment groups is needed) which is offset from the start of the required segments, i.e. at $T_1 + x$. In step 11.7, the media player 10 may identify using the manifest file that the time point $T_1 + x$
30 corresponds to a particular variant version, e.g. that which has a prior offset time closest to the switching time. In step 11.8, the identified variant version is requested from the content provider 30.

- 35 The content provider 30 at step 11.9 transmits the identified variant version of the or each required segment to the media player 10.

In step 11.10, the media player 10 receives the variant version of the or each required segment for buffering, decoding and rendering to the VR headset 20.

In summary, embodiments propose providing variate video streams at the server, or content provider 30 side to enable fast random access within segments. So, instead of placing the burden on the media player 10 to determine which part of a video stream segment need to be downloaded, in the present case the content provider 30 provides time-wise random access variants for each segment of video streams for the medial player 10 to choose from.

The methods and systems may be applied to MPEG-DASH standards. An additional data structure may be introduced for this purpose.

In order to guide the media player 10 to select the right variant for each situation, as mentioned above, a property may be provided in the manifest file 126 that indicates how to map a time offset to the corresponding variant URL. In normal cases, when the media player 10 continues decoding the same stream that it has decoded already, it uses the full segment, i.e. the first version. However, if it has to quickly switch to a video stream, and the playing time has already passed the start time of the segment, the media player 10 can select a variant that minimizes downloading video frames that are not needed any more.

In the disclosed embodiments, the burden placed on the server side can be kept minimal, because the variants can be generated by just packetizing the encoded stream in multiple ways, i.e. without a need to encode the same content with many parallel encoders. So, embodiments are also applicable for real-time applications, without a massive requirement for server HW capabilities.

As another embodiment of the invention, the server, or content provider 30, may also create separate normal streams and switching-optimized streams. For example, when a full segment is downloaded, the number of random access points can be kept minimal, but in the variants, the random access points can be inserted more often.

The embodiments still provide a seamless decoding experience, because one segment consists of full video Groups of Pictures (GOPs). In other words, the next segment after a variant segment starts with a full Intra frame, without any dependence on previous segments. Also, the segment time alignment (referred to in the DASH standard as the segment alignment attribute) may be kept, because the segmentStartOffset + segment variant duration must result in duration that is equal to the normal segment length.

It will be appreciated that the above described embodiments are purely illustrative and are not limiting on the scope of the invention. Other variations and modifications will be apparent to persons skilled in the art upon reading the present application.

- 20 -

Moreover, the disclosure of the present application should be understood to include any novel features or any novel combination of features either explicitly or implicitly disclosed herein or any generalization thereof and during the prosecution of the present application or of any application derived therefrom, new claims may be formulated to cover any such features and/or combination of such features.

5

Claims

1. A method comprising:

providing video data for streaming transmission to one or more remote devices, the video data being arranged into video segments representing respective temporal portions of video content, each video segment having a given start time and duration;

providing one or more variant versions of each segment in which the video data is divided into temporal sub-segments and in which one or more initial sub-segments contain substantially no video data and the other segment(s) contain video data from an offset time for the remainder of the duration;

receiving a request to transmit to a remote device either the video segment or a variant version; and

transmitting the requested video segment or variant version to the remote device.

2. The method of claim 1, wherein the video data is further arranged into groups of video segments, each group corresponding to a respective two-dimensional sub-area of a larger video image.

3. The method of claim 1 or claim 2, wherein the request is received responsive to a change in position of the remote device, or another device associated therewith, and the transmitting step comprises transmitting one or more new segments or variant versions corresponding to a new field-of-view.

4. The method of claim 3, wherein request is associated with a switching time offset from the start of a current segment, and the transmitting step comprises transmitting the variant version which has a prior offset time closest to the switching time.

5. The method of claim 4, further comprising, subsequent to transmitting one of the variant versions, transmitting the video segment corresponding to the next temporal portion in the event that no change in position is detected.

6. The method of any preceding claim, wherein the video data comprises a first variant version in which the first sub-segment contains substantially no video data, and one or more further variant versions in which an increasing number of successive sub-segment(s) contain substantially no video data.

7. The method of any preceding claim, further comprising providing a data file comprising, for each variant version, the associated offset time and means for identifying each variant version to enable downloading thereof from a remote device, and transmitting the data file to one or more remote devices.

8. The method of any preceding claim, wherein each video segment is divided into temporal sub-segments corresponding to those in each of the variant versions.

9. The method of any preceding claim, wherein each video segment comprises a first reference frame substantially at the start of the segment, and the variant version(s) each comprise different reference frames substantially at the start of each sub-segment which contains video data.

10. The method of claim 9, wherein the reference frame is an intra frame.

11. The method of claim 9 or claim 10, wherein each video segment and its variant versions further comprise one or more predicted frames based on the first reference frame, and wherein the variant versions further comprise the first reference frame to enable decoding of predicted frames in each sub-segment which contains video data.

12. The method of claim 11, wherein the first reference frame is provided substantially at the start of each variant version.

13. The method of any preceding claim, wherein the video data is virtual reality (VR) video data.

14. The method of any preceding claim, wherein the video segments are encoded using mono, stereo or multi-view coding.

15. The method of claim 13 or claim 14, wherein the request is received from a VR media player or VR headset.

16. The method of any preceding claim, wherein the request is a single HTTP request.

17. The method of claim 2, or any claim dependent thereon, wherein the video data represents a 360 degree video image.

18. A method comprising:

receiving streaming video data from a remote device for displaying a current field-of-view, the streaming data being received as a plurality of video segments representing respective temporal portions of video content, each video segment having a given start time and duration;

determining a new field-of-view; and

- 23 -

transmitting a request to the remote device for receiving a variant version of one or more segments corresponding to the new field-of-view, the variant version being divided into temporal sub-segments and in which one or more initial sub-segments contain substantially no video data and the other segment(s) contain video data from an offset time for the remainder of the duration.

19. The method of claim 18, wherein the video data is further arranged into groups of video segments, each group corresponding to a respective two-dimensional sub-area of a larger video image.

20. The method of claim 18 or claim 19, wherein the request is transmitted responsive to a change in position of a user device, or another device associated therewith.

21. The method of claim 20, further comprising receiving one or more new segments or variant versions corresponding to the new field-of-view.

22. The method of claim 20 or claim 21, wherein the new field-of-view is associated with a switching time which is offset from the start of a current segment, and the transmitting step comprises transmitting a request for the variant version which has a prior offset time closest to the switching time.

23. The method of claim 22, further comprising, subsequent to requesting one of the variant versions, transmitting a further request for the video segment corresponding to the next temporal portion in the event that no change in position is detected.

24. The method of any of claims 17 to 23, further comprising receiving a data file comprising, for each variant version of the segments, the associated offset time and means for identifying each variant version to enable requesting and receiving said variant version from the remote device.

25. The method of any of claims 17 to 24, wherein each received variant version(s) comprise reference frames substantially at the start of each sub-segment which contains video data.

26. The method of claim 25, wherein the reference frame is an intra frame.

27. The method of claim 25 or claim 26, wherein each received variant version(s) further comprise one or more predicted frames based on a first reference frame, and wherein the

received variant version(s) further comprise the first reference frame to enable decoding of predicted frames in each sub-segment which contains video data.

5 28. The method of claim 27, wherein the first reference frame is received substantially at the start of each variant version.

29. The method of any of claims 17 to 28, wherein the video data is virtual reality (VR) video data.

10 30. The method of claim 29, performed by VR media player or VR headset.

31. The method of any of claims 17 to 30, wherein the transmitting step comprises transmitting a single HTTP request.

15 32. The method of claim 18, or any claim dependent thereon, wherein the video data represents a 360 degree video image.

33. A computer program comprising instructions that when executed by a computer program control it to perform the method of any preceding claim.

20

34. A non-transitory computer-readable storage medium having stored thereon computer-readable code, which, when executed by at least one processor, causes the at least one processor to perform a method, comprising:

25 providing video data for streaming transmission to one or more remote devices, the video data being arranged into video segments representing respective temporal portions of video content, each video segment having a given start time and duration; providing one or more variant versions of each segment in which the video data is divided into temporal sub-segments and in which one or more initial sub-segments contain substantially no video data and the other segment(s) contain video data from
30 an offset time for the remainder of the duration; receiving a request to transmit to a remote device either the video segment or a variant version; and transmitting the requested video segment or variant version to the remote device.

35 35. An apparatus, the apparatus having at least one processor and at least one memory having computer-readable code stored thereon which when executed controls the at least one processor:

40 to provide video data for streaming transmission to one or more remote devices, the video data being arranged into video segments representing respective temporal portions of video content, each video segment having a given start time and duration;

- 25 -

to providing one or more variant versions of each segment in which the video data is divided into temporal sub-segments and in which one or more initial sub-segments contain substantially no video data and the other segment(s) contain video data from an offset time for the remainder of the duration;

5 to receiving a request to transmit to a remote device either the video segment or a variant version; and

to transmit the requested video segment or variant version to the remote device.

36. A non-transitory computer-readable storage medium having stored thereon
10 computer-readable code, which, when executed by at least one processor, causes the at least one processor to perform a method, comprising:

receiving streaming video data from a remote device for displaying a current field-of-view, the streaming data being received as a plurality of video segments representing respective temporal portions of video content, each video segment having a given start
15 time and duration;

determining a new field-of-view; and

transmitting a request to the remote device for receiving a variant version of one or more segments corresponding to the new field-of-view, the variant version being divided into temporal sub-segments and in which one or more initial sub-segments contain substantially
20 no video data and the other segment(s) contain video data from an offset time for the remainder of the duration.

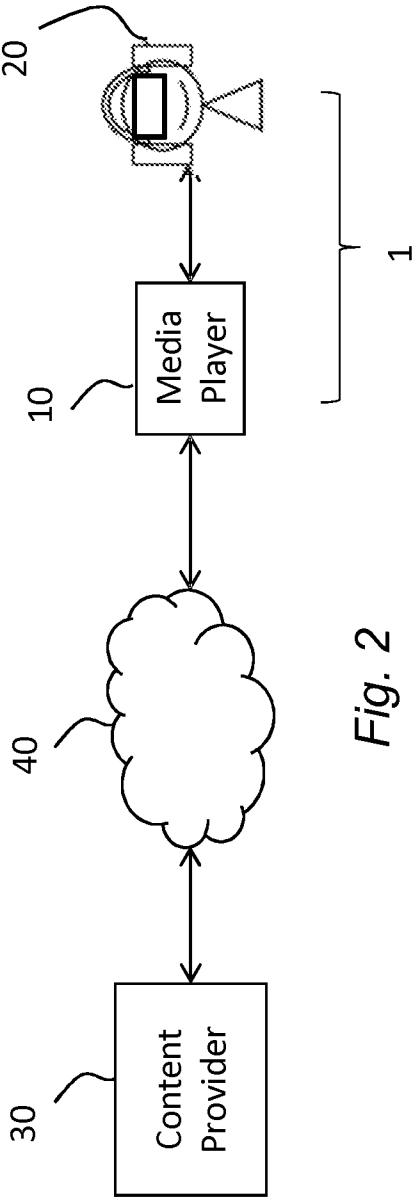
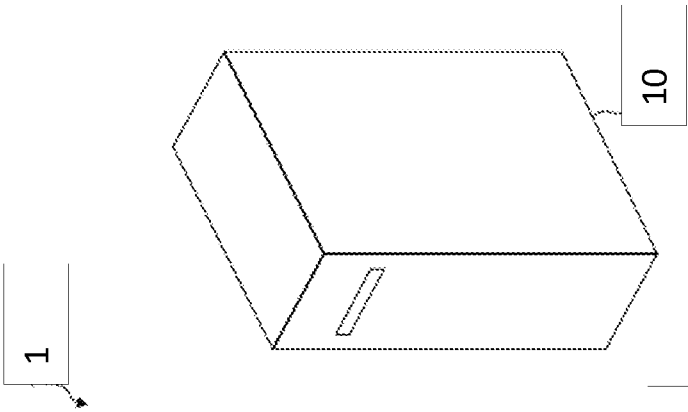
37. An apparatus, the apparatus having at least one processor and at least one memory having computer-readable code stored thereon which when executed controls the at least one
25 processor:

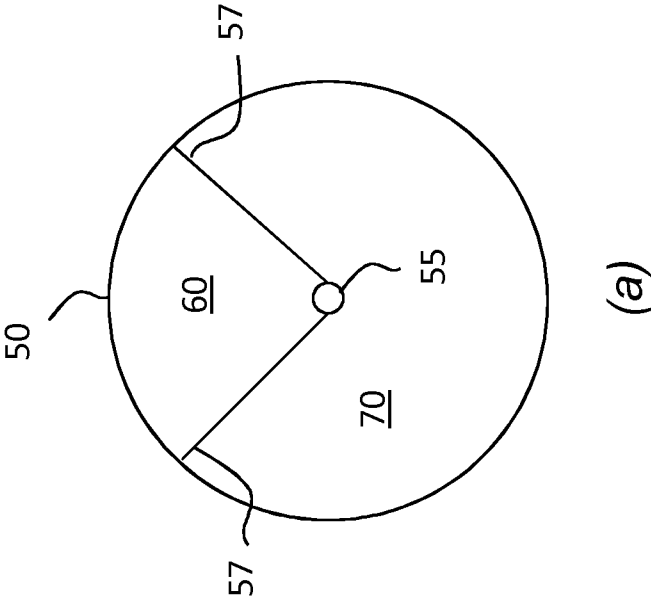
to receive streaming video data from a remote device for displaying a current field-of-view, the streaming data being received as a plurality of video segments representing respective temporal portions of video content, each video segment having a given start time and duration;

30 to determining a new field-of-view; and

to transmit a request to the remote device for receiving a variant version of one or more segments corresponding to the new field-of-view, the variant version being divided into temporal sub-segments and in which one or more initial sub-segments contain substantially no video data and the other segment(s) contain video data from an offset
35 time for the remainder of the duration.

38. An apparatus configured to perform the method of any of claims 1 to 32.





	<u>80a</u>	<u>80b</u>	<u>80c</u>	<u>80d</u>	
	<u>80e</u>	<u>80f</u>	<u>80g</u>	<u>80h</u>	

(b)

Fig. 3

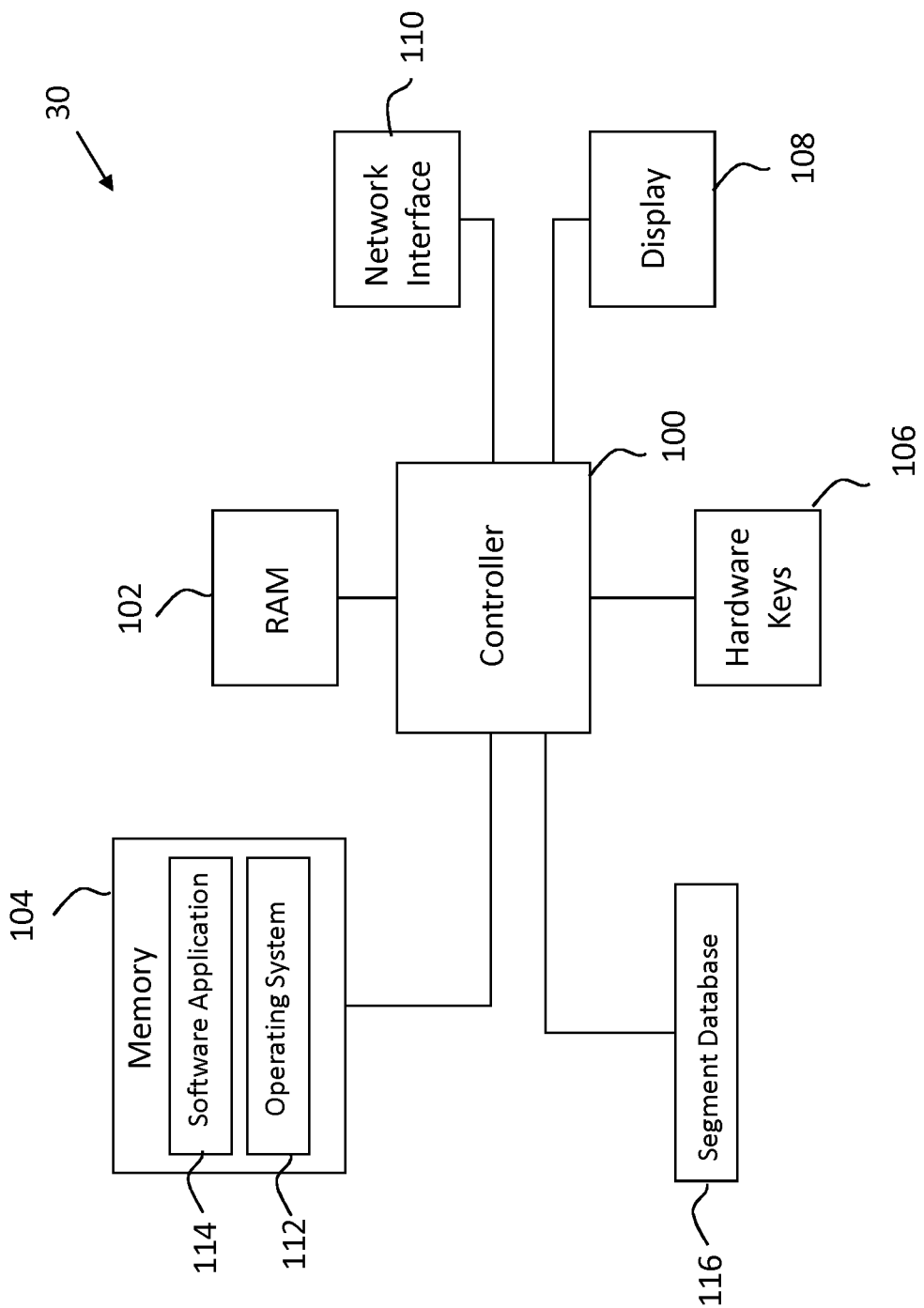


Fig. 4

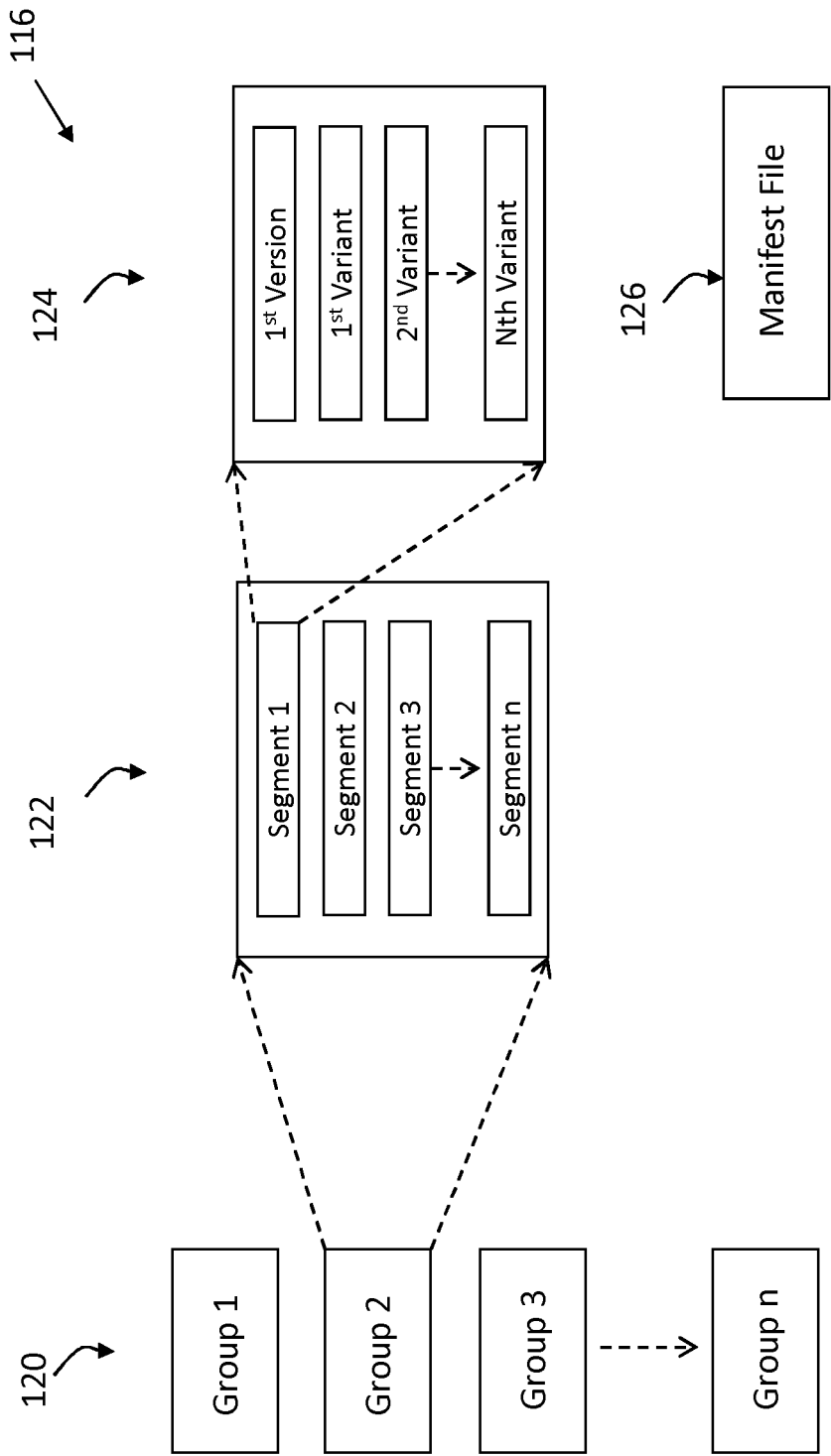


Fig. 5

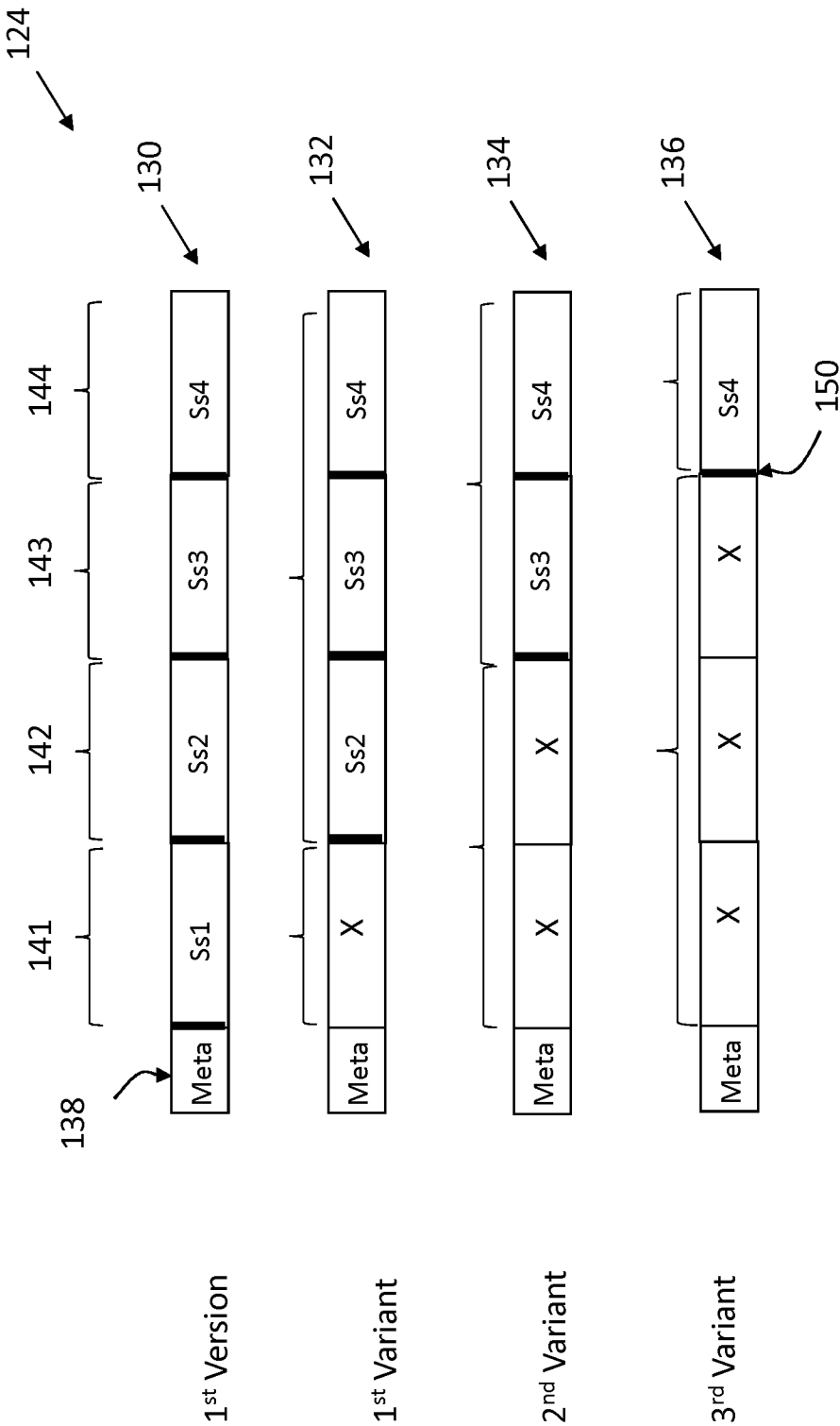


Fig. 6

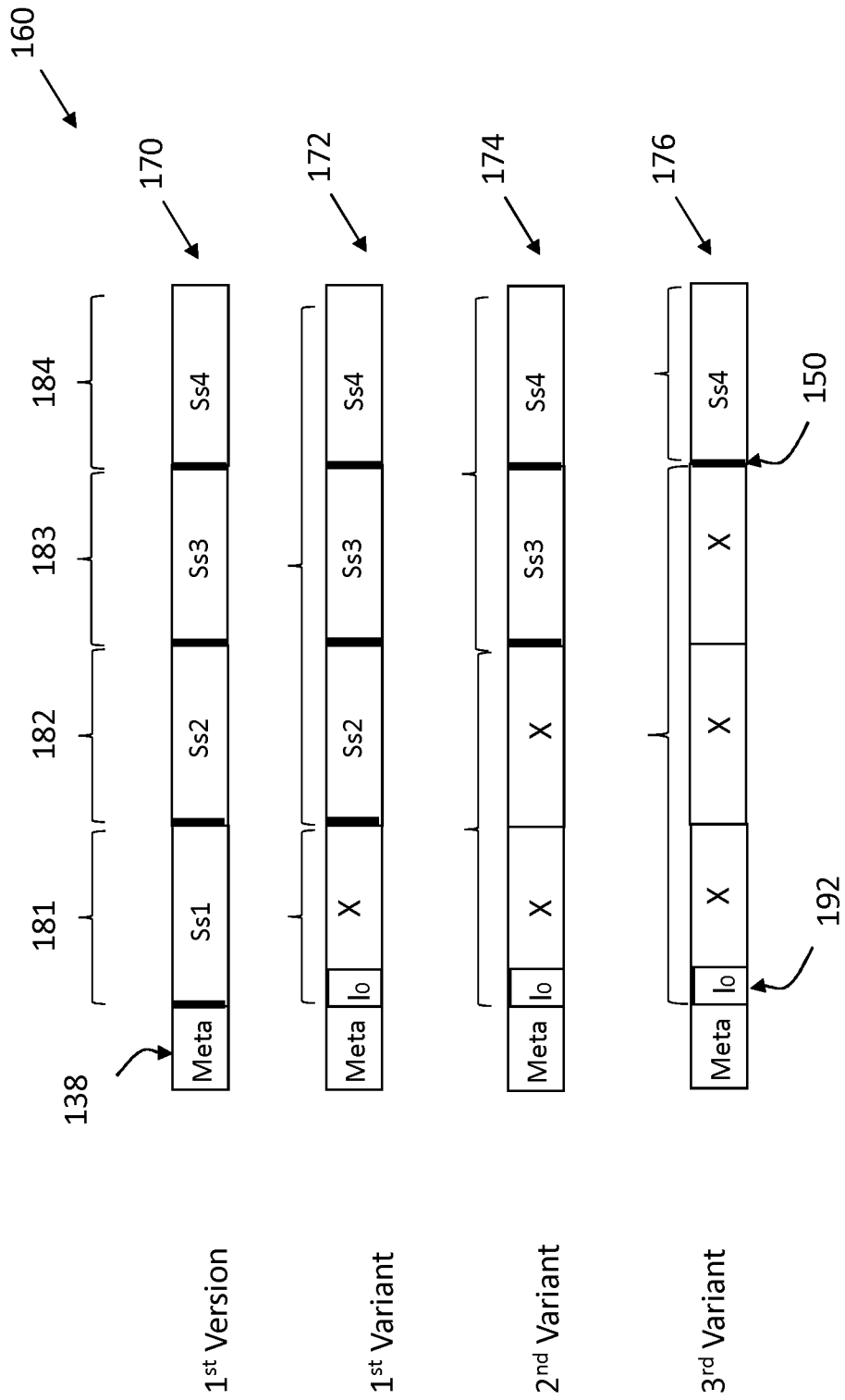


Fig. 7

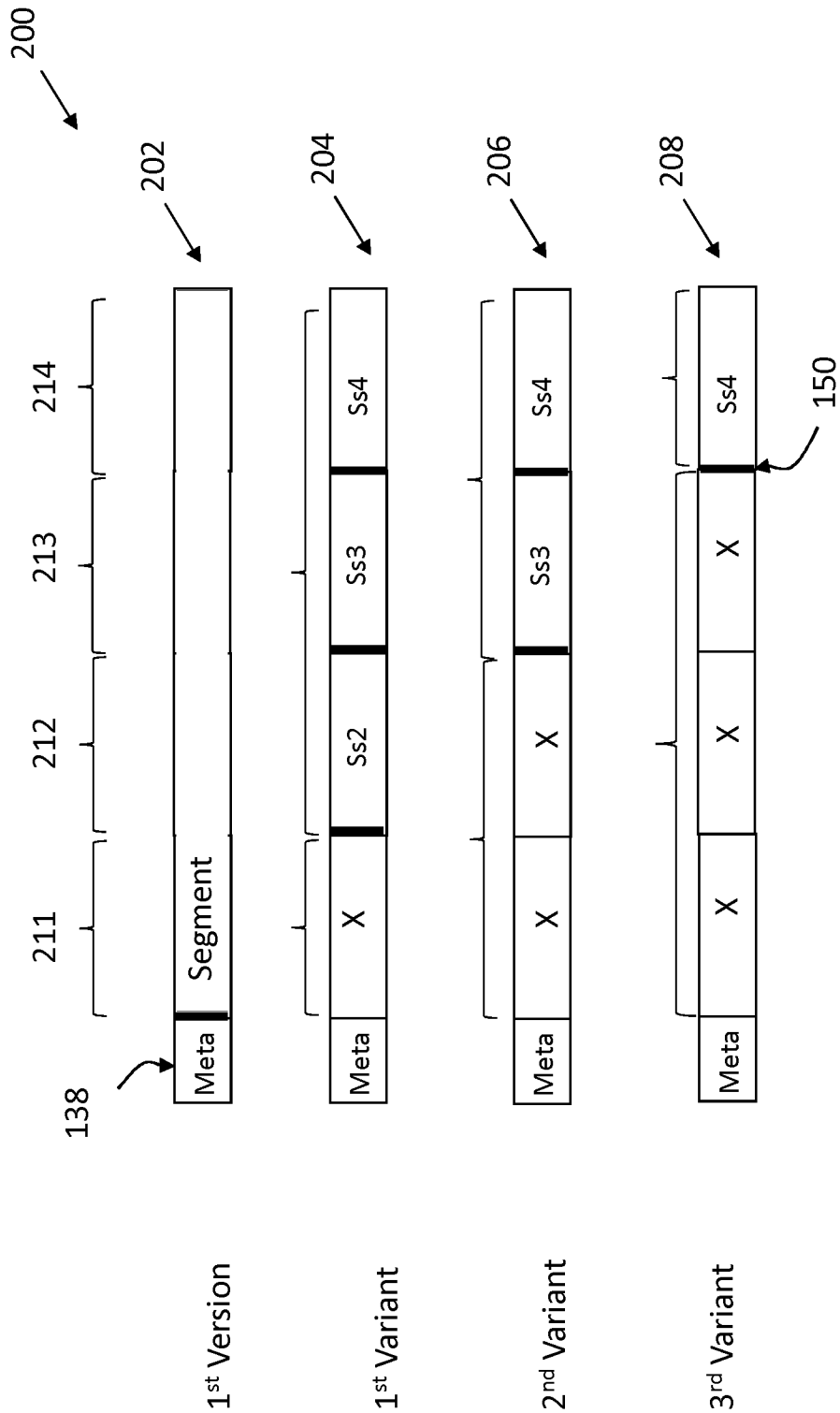


Fig. 8

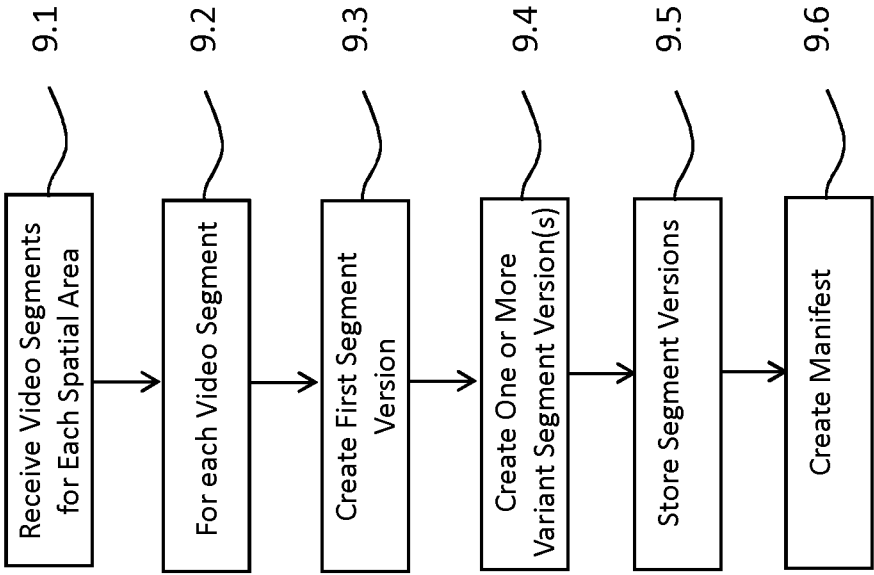


Fig. 9

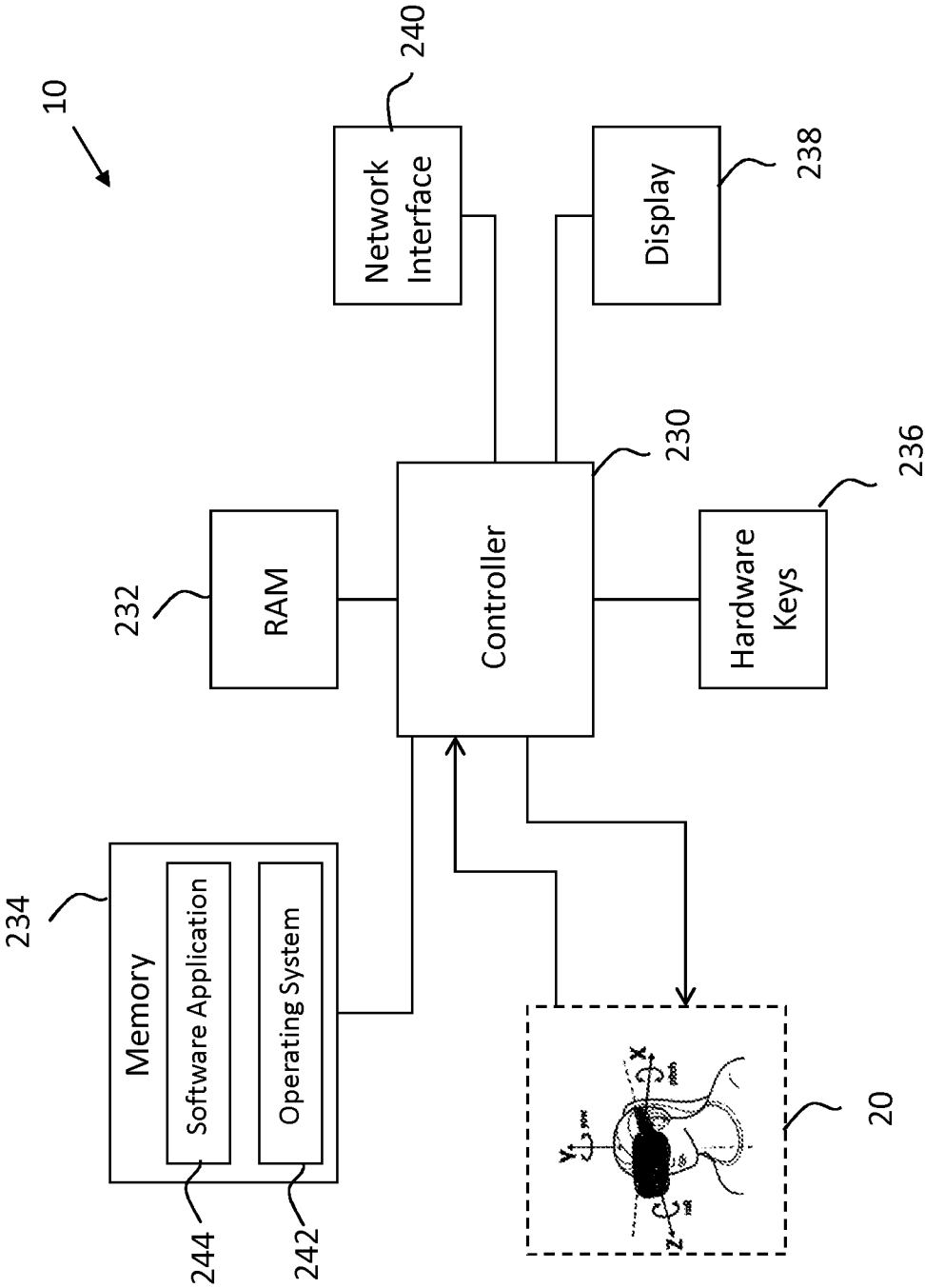


Fig. 10

Media Player 10

Content Provider 30

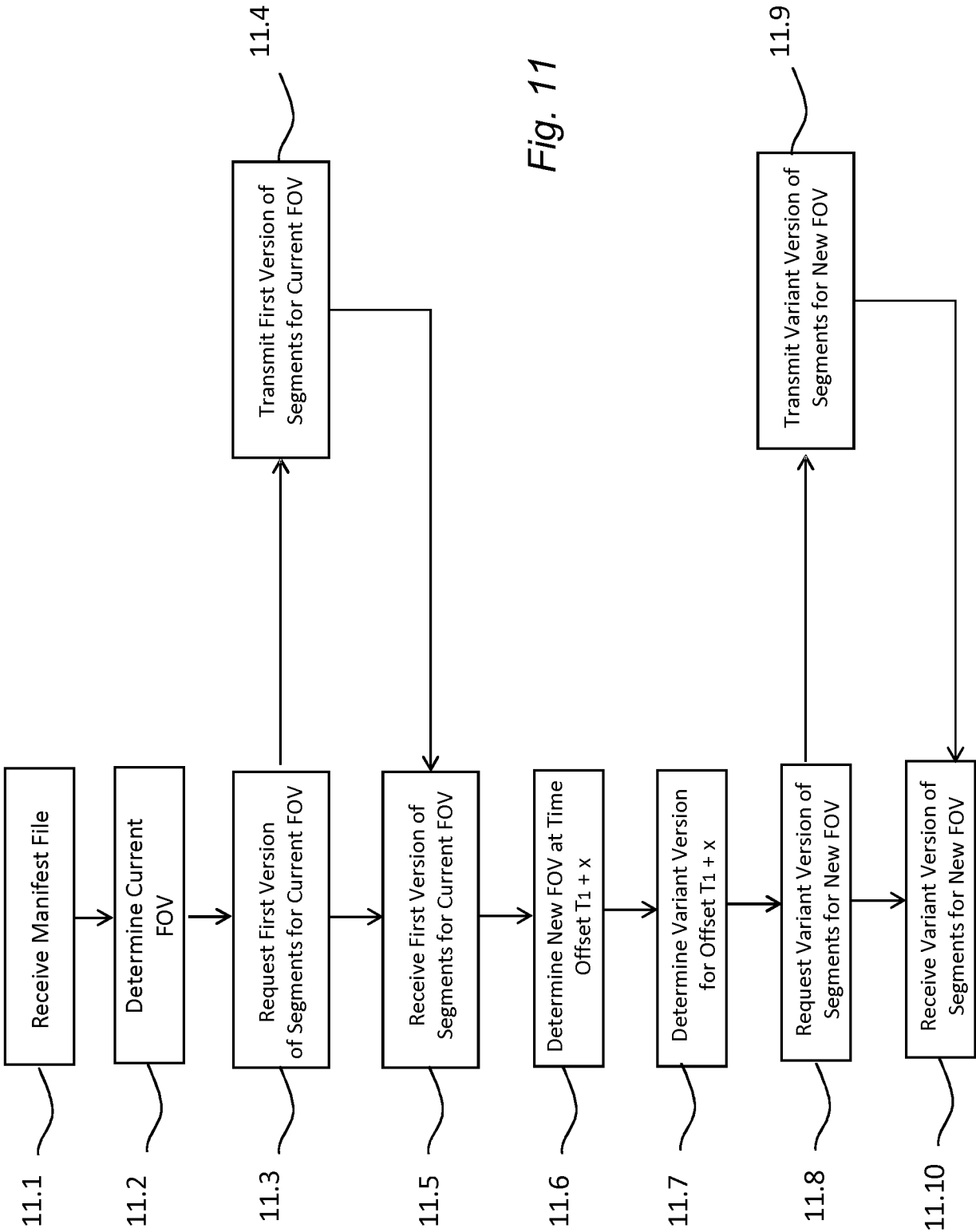


Fig. 11