(12) UK Patent (19) GB (11) 2597725 (13) B

(45) Date of B Publication 21.02.2024

(54) Title of the Invention: Data processing system and method for image enhancement

(51) INT CL: ***G06T 3/40*** (2024.01) ***G06T 5/00*** (2024.01) *G06T 19/00* (2011.01)

(21) Application No: **2011931.9**

(22) Date of Filing: **31.07.2020**

(43) Date of A Publication **09.02.2022**

(56) Documents Cited:
US 20190026874 A1 **US 20190026864 A1**
US 20170287446 A1
ACM TRANSACTIONS ON GRAPHICS, vol 35, 2016, ANJUL PATNEY ET AL, "Towards foveated rendering for gaze-tracked virtual reality" pages 1-12

(58) Field of Search:
As for published application 2597725 A viz:
INT CL **G06T**
updated as appropriate

Additional Fields
Other: **None**

(72) Inventor(s):
Fabio Cappello
Maria Chiara Monti
Alexander Edward James Smith
Rajeev Gupta
Alexei Ashton Derek Smith
Patrick John Connor

(73) Proprietor(s):
Sony Interactive Entertainment Inc.
1-7-1 Konan, Minato-Ku 108-0075, Tokyo, Japan

(74) Agent and/or Address for Service:
D Young & Co LLP
120 Holborn, LONDON, EC1N 2DY, United Kingdom

GB 2597725 B

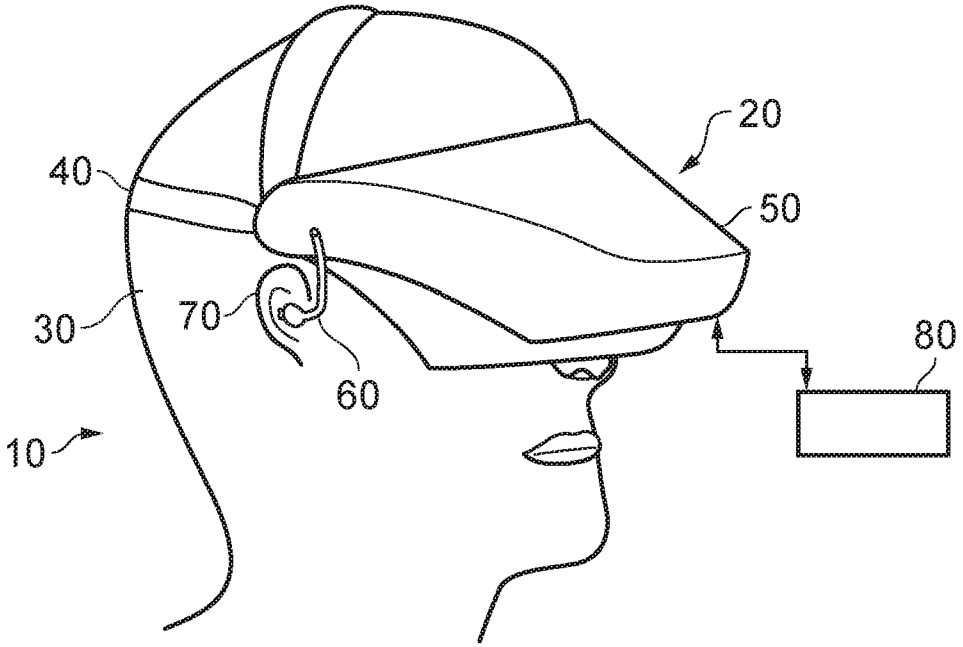
20
40
50
30
70
80
60
10

FIG. 1

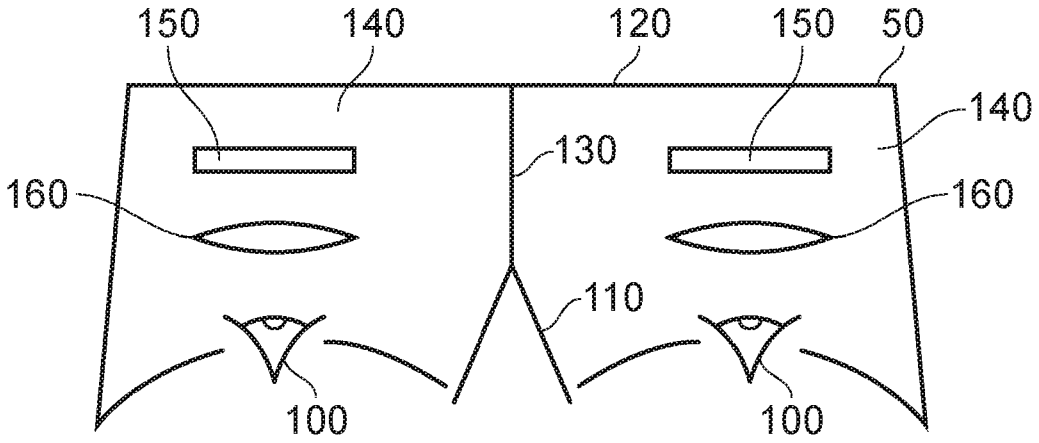
150
140
120
150
50
140
130
160
160
110
100
100

FIG. 2

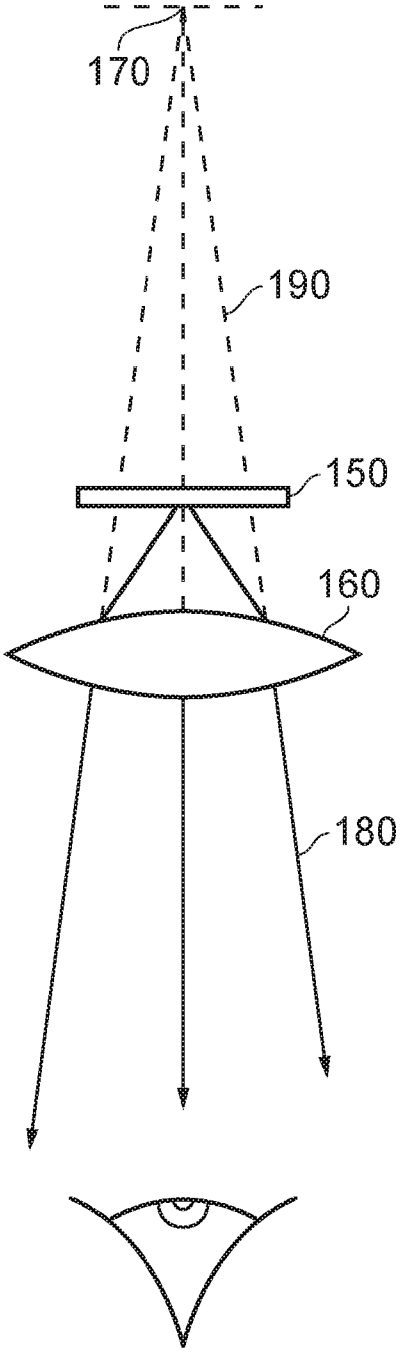

FIG. 3

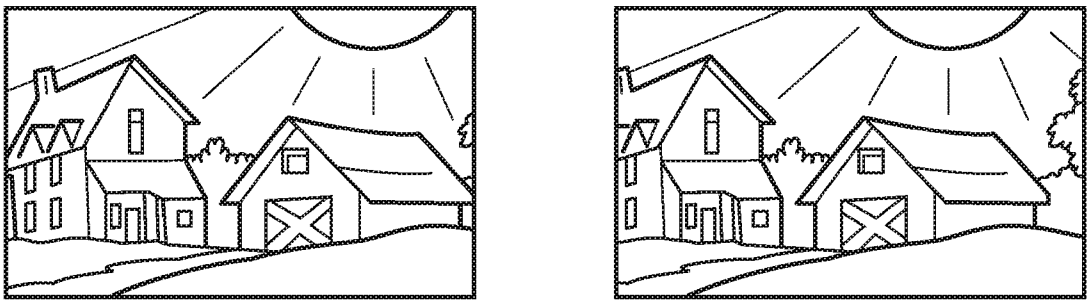

FIG. 5

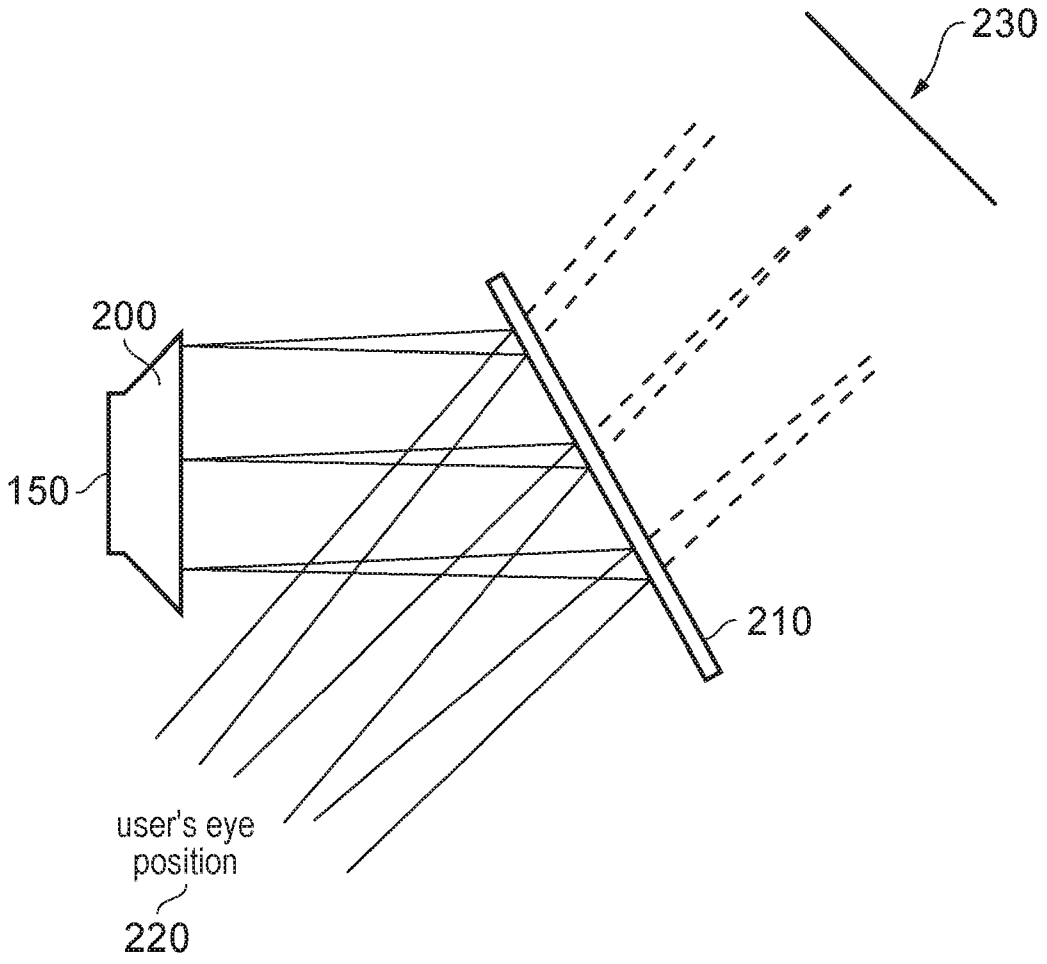

FIG. 4

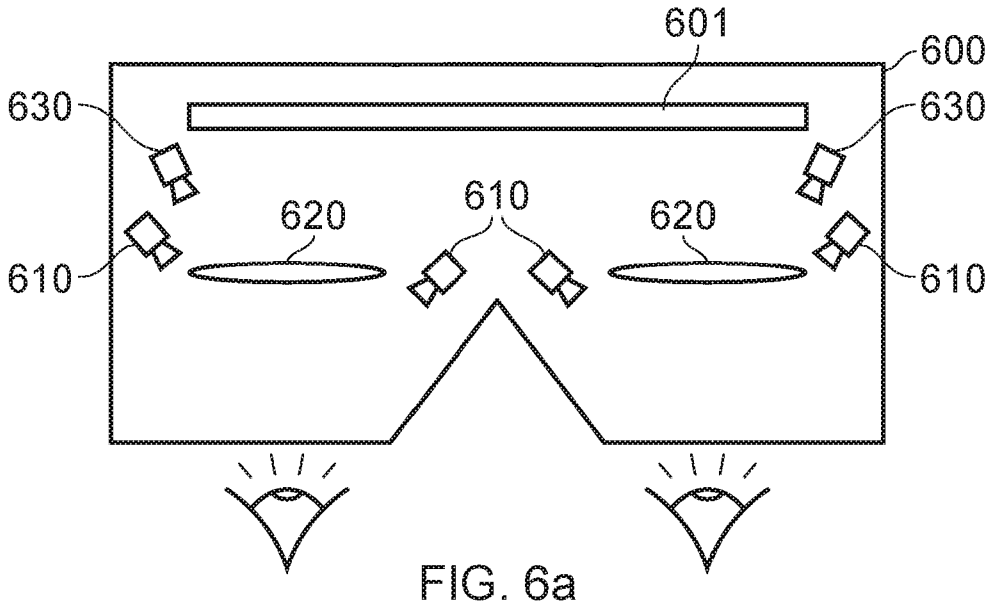

FIG. 6a

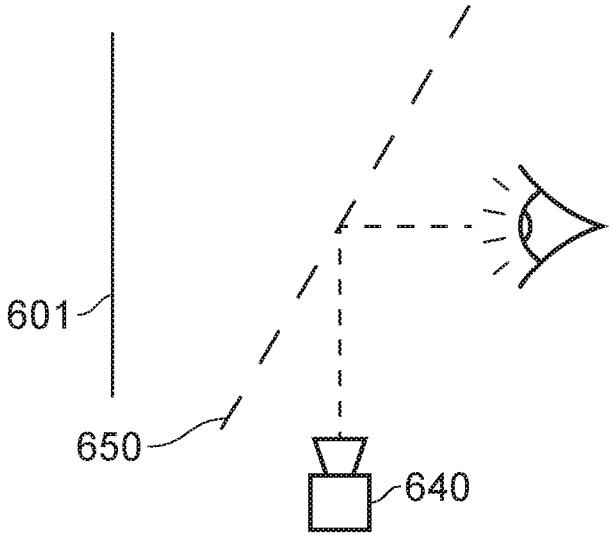

FIG. 6b

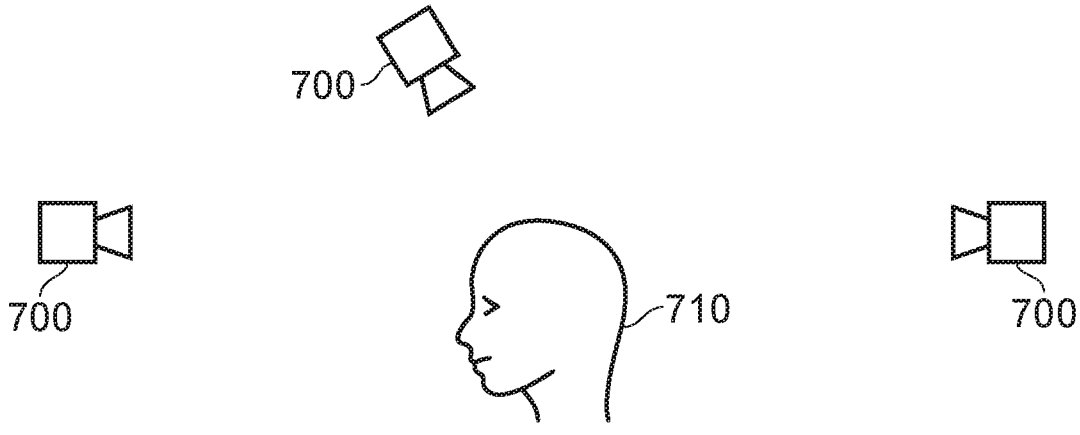

FIG. 7

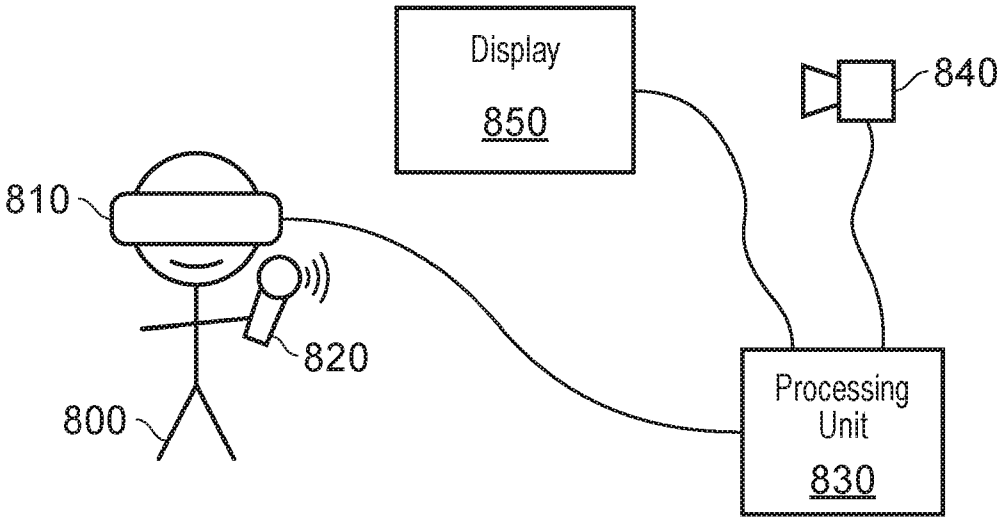
Display
850
840
810
820
800
Processing
Unit
830

FIG. 8

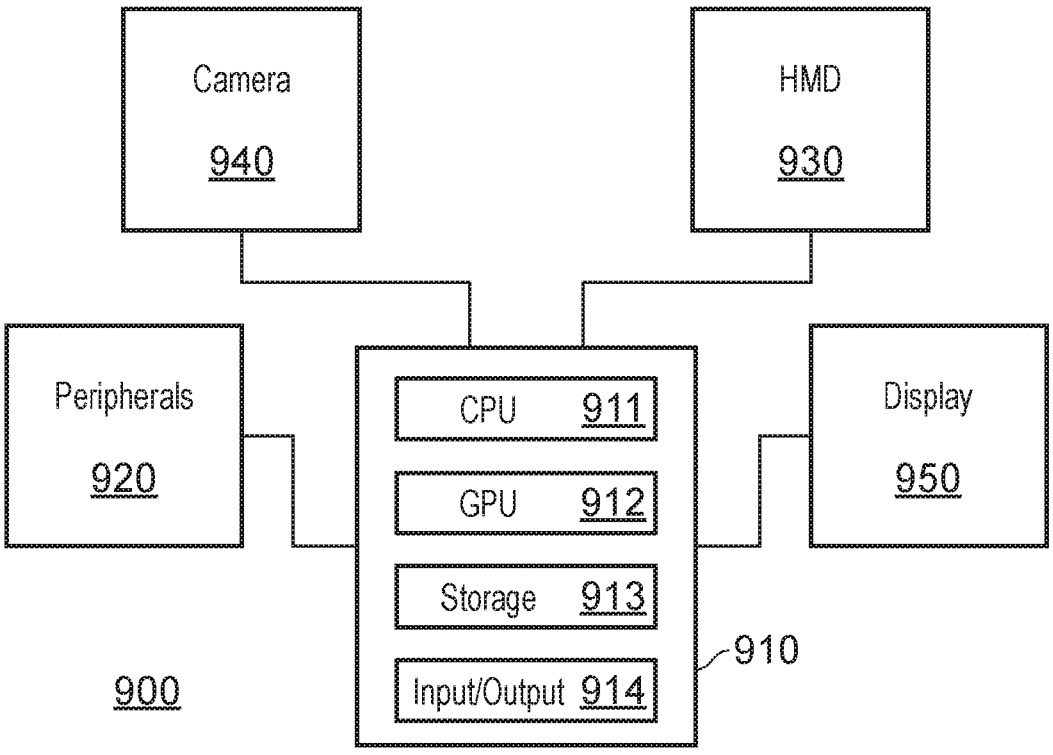
Camera
940
HMD
930
Peripherals
920
CPU 911
GPU 912
Storage 913
Input/Output 914
910
Display
950
900

FIG. 9

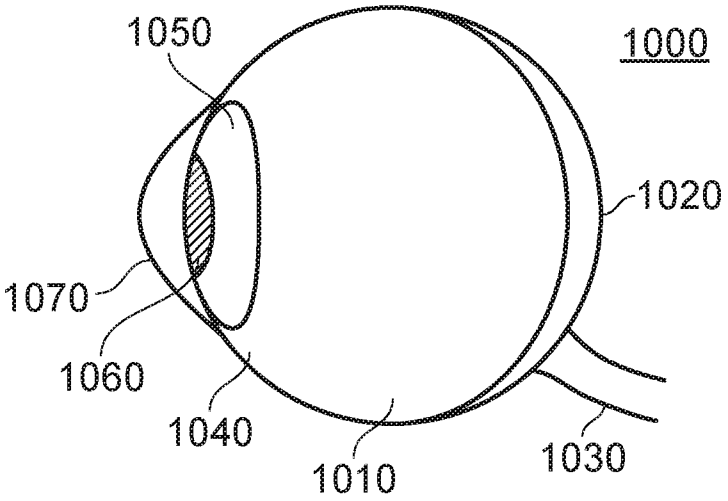
1050
1000
1020
1070
1060
1040
1010
1030

FIG. 10

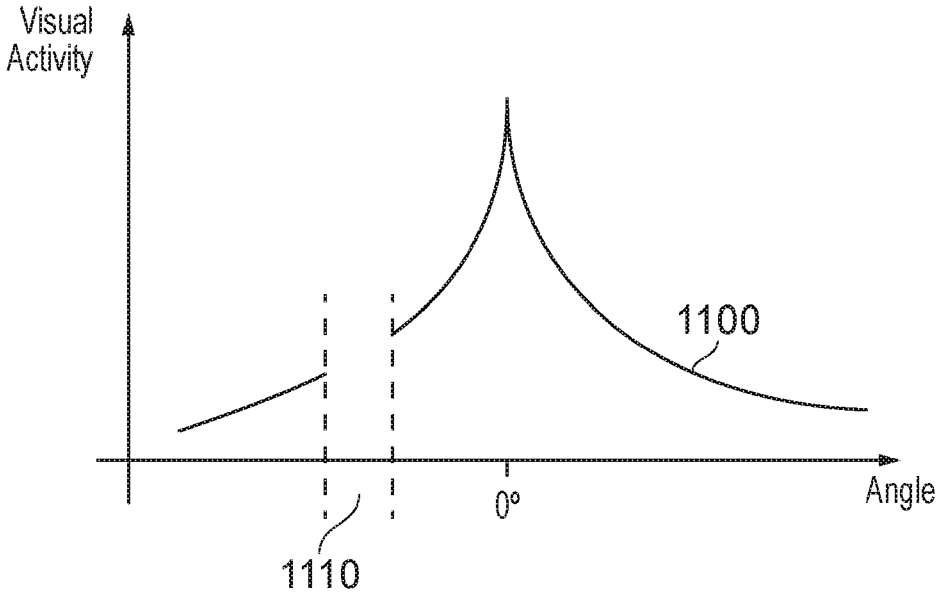
Visual
Activity
1100
0°
Angle
1110

FIG. 11

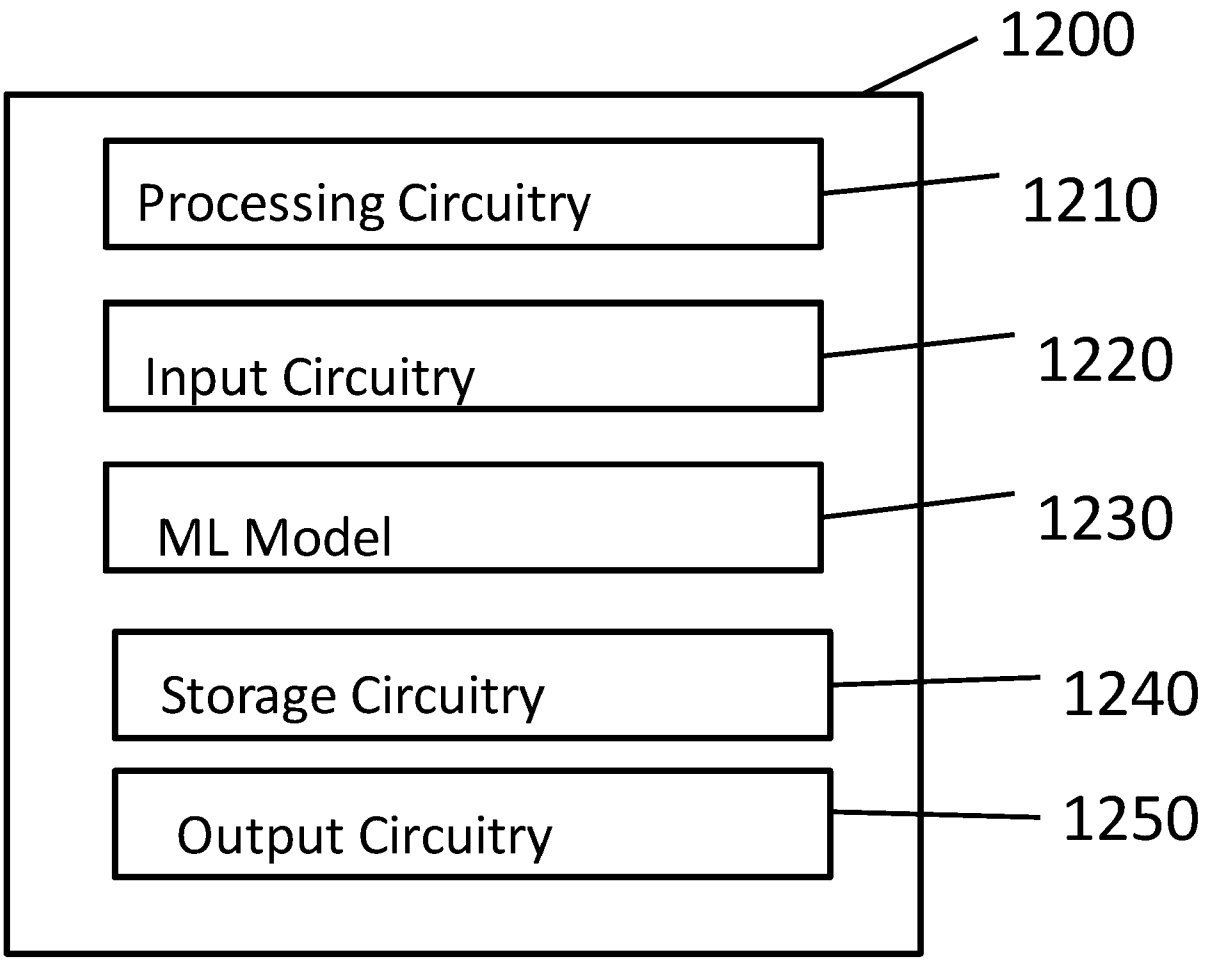

Fig. 12

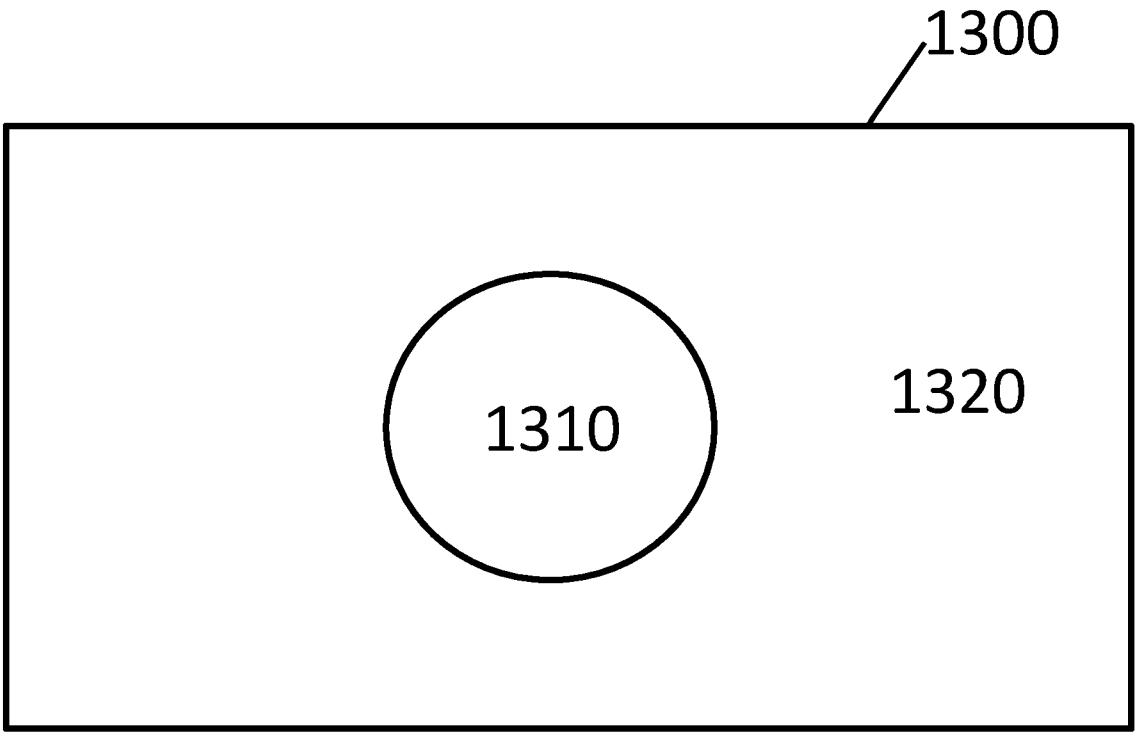

Fig. 13a

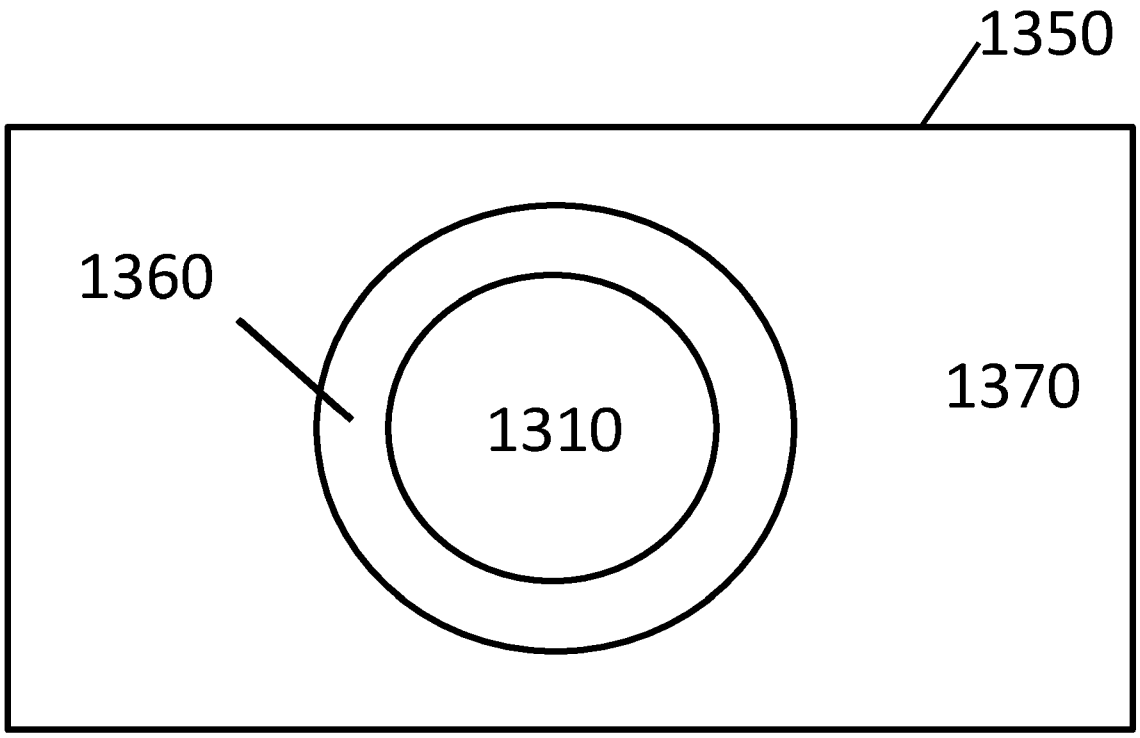

Fig. 13b

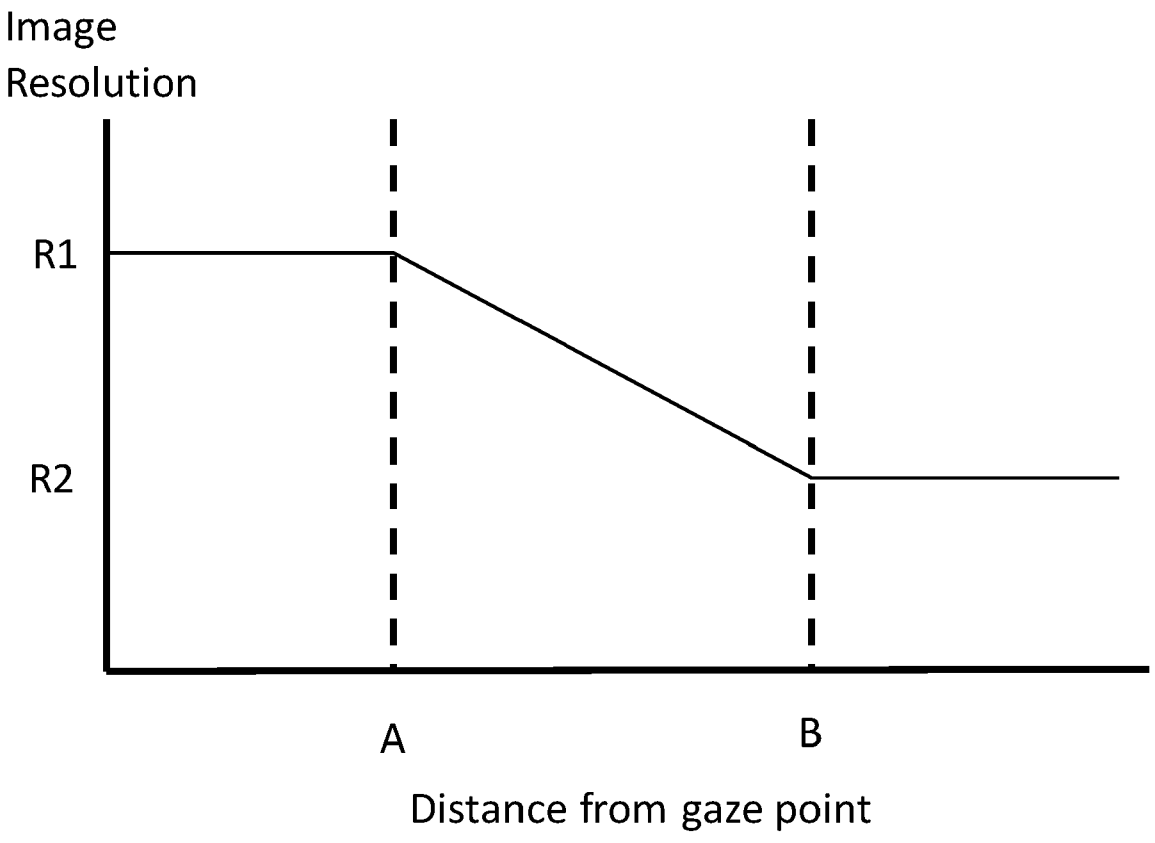

Fig. 14a

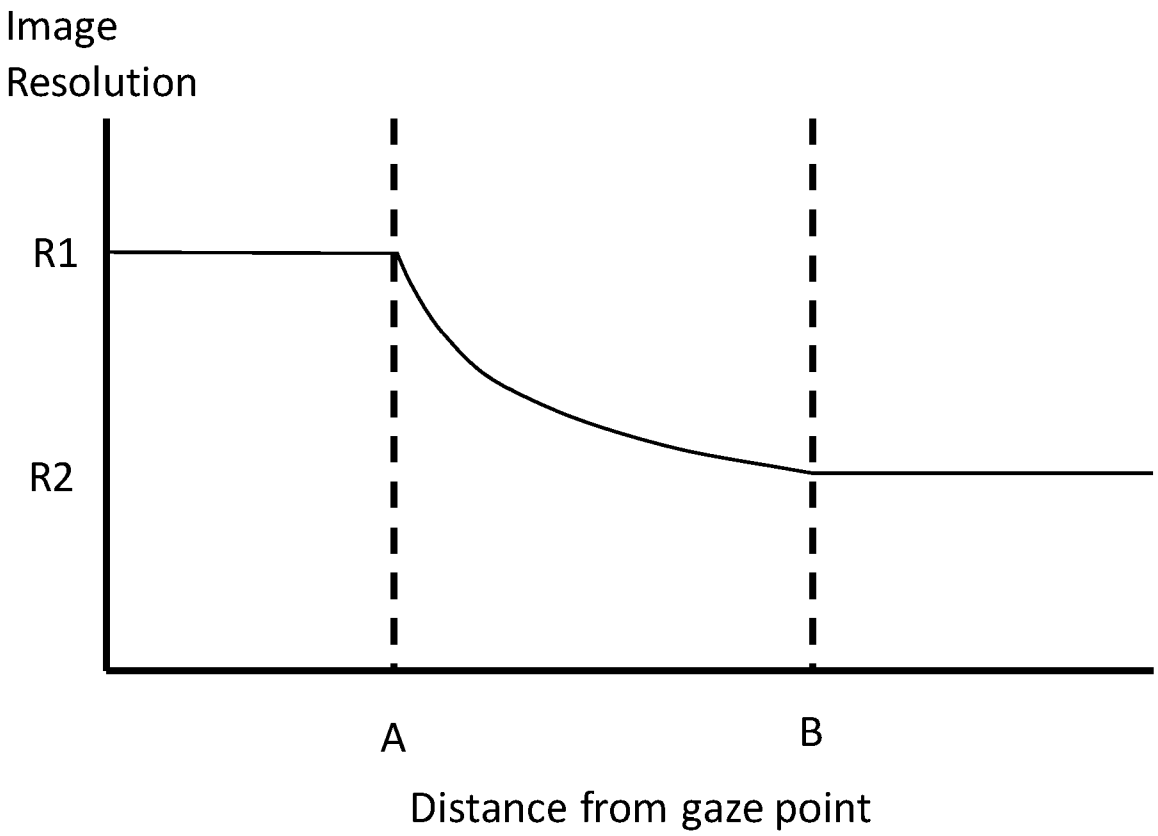

Fig. 14b

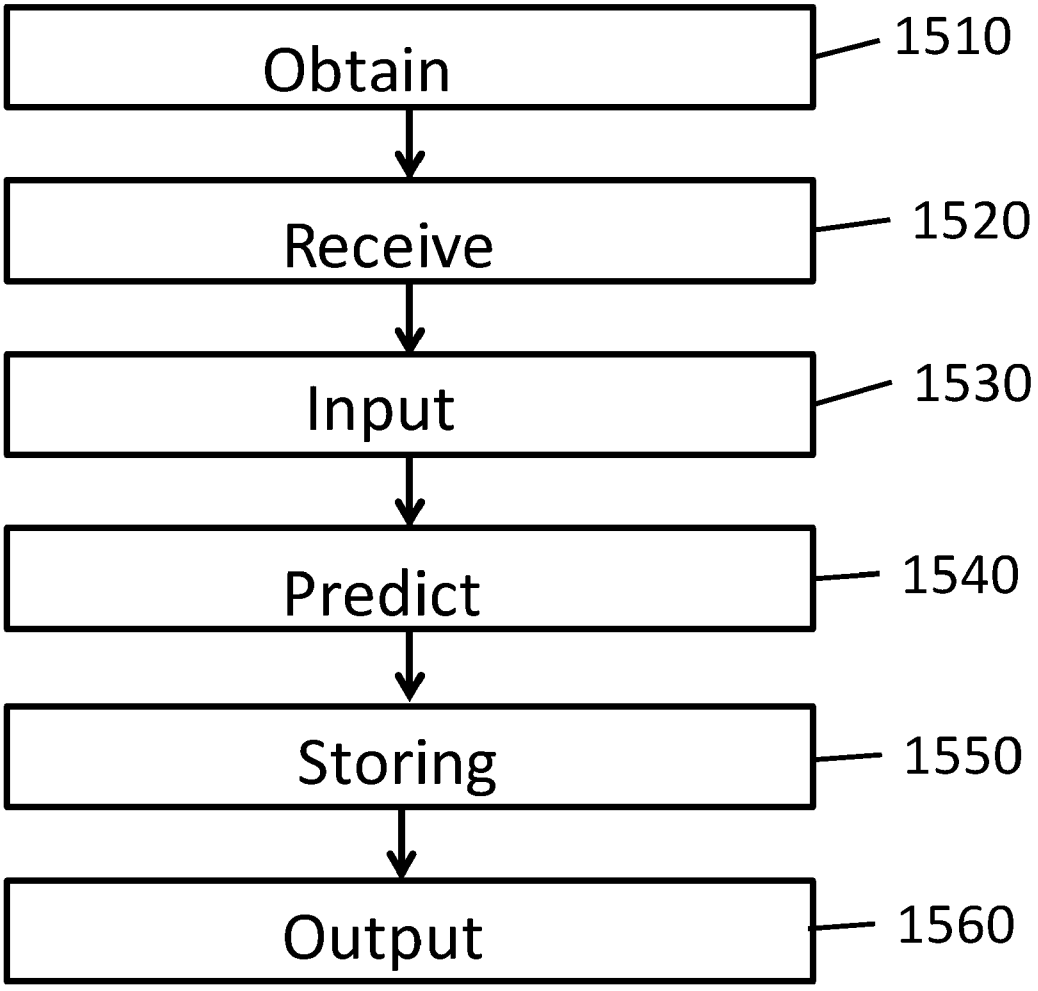
Obtain
1510
Receive
1520
Input
1530
Predict
1540
Storing
1550
Output
1560

Fig. 15

DATA PROCESSING SYSTEM AND METHOD FOR IMAGE ENHANCEMENT

The present disclosure relates to data processing systems and methods for image enhancement. In particular, the present disclosure relates to data processing systems and methods that use gaze data from gaze tracking systems and pixel values from image frames to obtain additional pixel values for enhancing the image frames.

Gaze tracking systems are used to identify a location of a subject's gaze within an environment; in many cases, this location may be a position on a display screen that is being viewed by the subject. In a number of existing arrangements, this is performed using one or more inwards-facing cameras directed towards the subject's eye (or eyes) in order to determine a direction in which the eyes are oriented at any given time. Having identified the orientation of the eye, a gaze direction can be determined and a focal region may be determined as the intersection of the gaze direction of each eye.

One application for which gaze tracking is considered of particular use is that of use in head-mountable display units (HMDs). The use in HMDs may be of particular benefit owing to the close proximity of inward-facing cameras to the user's eyes, allowing the tracking to be performed much more accurately and precisely than in arrangements in which it is not possibly to provide the cameras with such proximity.

By utilising gaze detection techniques, it may be possible to provide a more efficient and/or effective processing method for generating content or interacting with devices.

For example, gaze tracking may be used to provide user inputs or to assist with such inputs – a continued gaze at a location may act as a selection, or a gaze towards a particular object accompanied by another input (such as a button press) may be considered as a suitable input. This may be more effective as an input method in some embodiments, particularly in those in which a controller is not provided or when a user has limited mobility.

Foveal rendering is an example of a use for the results of a gaze tracking process in order to improve the efficiency of a content generation process. Foveal rendering is rendering that is performed so as to exploit the fact that human vision is only able to identify high detail in a narrow region (the fovea), with the ability to discern detail tailing off sharply outside of this region.

In such methods, a portion of the display can be identified as being an area of focus in accordance with the user's gaze direction. This portion of the display can be supplied with high-quality image content, while the remaining areas of the display can be provided with lower-quality (and therefore less resource intensive to generate) image content. This can lead to a more efficient use of available processing resources without a noticeable degradation of image quality for the user.

It is therefore considered advantageous to be able to improve gaze tracking methods, and/or apply the results of such methods in an improved manner. It is in the context of such advantages that the present disclosure arises.

Various aspects and features of the present invention are defined in the appended claims and within the text of the accompanying description.

Embodiments of the invention will now be described, by way of example only, with reference to the accompanying drawings, in which:

Figure 1 schematically illustrates an HMD worn by a user;

Figure 2 is a schematic plan view of an HMD;

Figure 3 schematically illustrates the formation of a virtual image by an HMD;

Figure 4 schematically illustrates another type of display for use in an HMD;

Figure 5 schematically illustrates a pair of stereoscopic images;

Figure 6a schematically illustrates a plan view of an HMD;

Figure 6b schematically illustrates a near-eye tracking arrangement;

Figure 7 schematically illustrates a remote tracking arrangement;

Figure 8 schematically illustrates a gaze tracking environment;

Figure 9 schematically illustrates a gaze tracking system;

Figure 10 schematically illustrates a human eye;

Figure 11 schematically illustrates a graph of human visual acuity;

Figure 12 schematically illustrates a data processing apparatus;

Figure 13a schematically illustrates an example of a predicted image frame;

Figure 13b schematically illustrates an example of another predicted image frame;

Figure 14a schematically illustrates a graph of image resolution versus distance from a gaze point;

Figure 14b schematically illustrates another graph of image resolution versus distance from a gaze point; and

Figure 15 is a schematic flowchart illustrating a data processing method.

Referring now to Figure 1, a user 10 is wearing an HMD 20 (as an example of a generic head-mountable apparatus – other examples including audio headphones or a head-mountable light source) on the user's head 30. The HMD comprises a frame 40, in this example formed of a rear strap and a top strap, and a display portion 50. As noted above, many gaze tracking arrangements may be considered particularly suitable for use in HMD systems; however, use with such an HMD system should not be considered essential.

Note that the HMD of Figure 1 may comprise further features, to be described below in connection with other drawings, but which are not shown in Figure 1 for clarity of this initial explanation.

The HMD of Figure 1 completely (or at least substantially completely) obscures the user's view of the surrounding environment. All that the user can see is the pair of images displayed within the HMD, as supplied by an external processing device such as a games console in many embodiments. Of course, in some embodiments images may instead (or additionally) be generated by a processor or obtained from memory located at the HMD itself.

The HMD has associated headphone audio transducers or earpieces 60 which fit into the user's left and right ears 70. The earpieces 60 replay an audio signal provided from an external source, which may be the same as the video signal source which provides the video signal for display to the user's eyes.

The combination of the fact that the user can see only what is displayed by the HMD and, subject to the limitations of the noise blocking or active cancellation properties of the earpieces and associated electronics, can hear only what is provided via the earpieces, mean that this HMD may be considered as a so-called “full immersion” HMD. Note however that in some embodiments the HMD is not a full immersion HMD, and may provide at least some facility for the user to see and/or hear the user’s surroundings. This could be by providing some degree of transparency or partial transparency in the display arrangements, and/or by projecting a view of the outside (captured using a camera, for example a camera mounted on the HMD) via the HMD’s displays, and/or by allowing the transmission of ambient sound past the earpieces and/or by providing a microphone to generate an input sound signal (for transmission to the earpieces) dependent upon the ambient sound.

A front-facing camera 122 may capture images to the front of the HMD, in use. Such images may be used for head tracking purposes, in some embodiments, while it may also be suitable for capturing images for an augmented reality (AR) style experience. A Bluetooth® antenna 124 may provide communication facilities or may simply be arranged as a directional antenna to allow a detection of the direction of a nearby Bluetooth® transmitter.

In operation, a video signal is provided for display by the HMD. This could be provided by an external video signal source 80 such as a video games machine or data processing apparatus (such as a personal computer), in which case the signals could be transmitted to the HMD by a wired or a wireless connection 82. Examples of suitable wireless connections include Bluetooth® connections. Audio signals for the earpieces 60 can be carried by the same connection. Similarly, any control signals passed from the HMD to the video (audio) signal source may be carried by the same connection. Furthermore, a power supply 83 (including one or more batteries and/or being connectable to a mains power outlet) may be linked by a cable 84 to the HMD. Note that the power supply 83 and the video signal source 80 may be separate units or may be embodied as the same physical unit. There may be separate cables for power and video (and indeed for audio) signal supply, or these may be combined for carriage on a single cable (for example, using separate conductors, as in a USB cable, or in a similar way to a

"power over Ethernet" arrangement in which data is carried as a balanced signal and power as direct current, over the same collection of physical wires). The video and/or audio signal may be carried by, for example, an optical fibre cable. In other embodiments, at least part of the functionality associated with generating image and/or audio signals for presentation to the user may be carried out by circuitry and/or processing forming part of the HMD itself. A power supply may be provided as part of the HMD itself.

Some embodiments of the invention are applicable to an HMD having at least one electrical and/or optical cable linking the HMD to another device, such as a power supply and/or a video (and/or audio) signal source. So, embodiments of the invention can include, for example:

(a) an HMD having its own power supply (as part of the HMD arrangement) but a cabled connection to a video and/or audio signal source;

(b) an HMD having a cabled connection to a power supply and to a video and/or audio signal source, embodied as a single physical cable or more than one physical cable;

(c) an HMD having its own video and/or audio signal source (as part of the HMD arrangement) and a cabled connection to a power supply; or

(d) an HMD having a wireless connection to a video and/or audio signal source and a cabled connection to a power supply.

If one or more cables are used, the physical position at which the cable 82 and/or 84 enters or joins the HMD is not particularly important from a technical point of view. Aesthetically, and to avoid the cable(s) brushing the user's face in operation, it would normally be the case that the cable(s) would enter or join the HMD at the side or back of the HMD (relative to the orientation of the user's head when worn in normal operation). Accordingly, the position of the cables 82, 84 relative to the HMD in Figure 1 should be treated merely as a schematic representation.

Accordingly, the arrangement of Figure 1 provides an example of a head-mountable display system comprising a frame to be mounted onto an observer's head, the frame defining one or two eye display positions which, in use, are positioned in front of a respective eye of the observer and a display element mounted with respect to each of the eye display positions, the display element providing a virtual image of a video display of a video signal from a video signal source to that eye of the observer.

Figure 1 shows just one example of an HMD. Other formats are possible: for example an HMD could use a frame more similar to that associated with conventional eyeglasses, namely a substantially horizontal leg extending back from the display portion to the top rear of the user's ear, possibly curling down behind the ear. In other (not full immersion) examples, the user's view of the external environment may not in fact be entirely obscured; the displayed images could be arranged so as to be superposed (from the user's point of view) over the

external environment. An example of such an arrangement will be described below with reference to Figure 4.

In the example of Figure 1, a separate respective display is provided for each of the user's eyes. A schematic plan view of how this is achieved is provided as Figure 2, which illustrates the positions 100 of the user's eyes and the relative position 110 of the user's nose. The display portion 50, in schematic form, comprises an exterior shield 120 to mask ambient light from the user's eyes and an internal shield 130 which prevents one eye from seeing the display intended for the other eye. The combination of the user's face, the exterior shield 120 and the interior shield 130 form two compartments 140, one for each eye. In each of the compartments there is provided a display element 150 and one or more optical elements 160. The way in which the display element and the optical element(s) cooperate to provide a display to the user will be described with reference to Figure 3.

Referring to Figure 3, the display element 150 generates a displayed image which is (in this example) refracted by the optical elements 160 (shown schematically as a convex lens but which could include compound lenses or other elements) so as to generate a virtual image 170 which appears to the user to be larger than and significantly further away than the real image generated by the display element 150. As an example, the virtual image may have an apparent image size (image diagonal) of more than 1 m and may be disposed at a distance of more than 1 m from the user's eye (or from the frame of the HMD). In general terms, depending on the purpose of the HMD, it is desirable to have the virtual image disposed a significant distance from the user. For example, if the HMD is for viewing movies or the like, it is desirable that the user's eyes are relaxed during such viewing, which requires a distance (to the virtual image) of at least several metres. In Figure 3, solid lines (such as the line 180) are used to denote real optical rays, whereas broken lines (such as the line 190) are used to denote virtual rays.

An alternative arrangement is shown in Figure 4. This arrangement may be used where it is desired that the user's view of the external environment is not entirely obscured. However, it is also applicable to HMDs in which the user's external view is wholly obscured. In the arrangement of Figure 4, the display element 150 and optical elements 200 cooperate to provide an image which is projected onto a mirror 210, which deflects the image towards the user's eye position 220. The user perceives a virtual image to be located at a position 230 which is in front of the user and at a suitable distance from the user.

In the case of an HMD in which the user's view of the external surroundings is entirely obscured, the mirror 210 can be a substantially 100% reflective mirror. The arrangement of Figure 4 then has the advantage that the display element and optical elements can be located closer to the centre of gravity of the user's head and to the side of the user's eyes, which can produce a less bulky HMD for the user to wear. Alternatively, if the HMD is designed not to completely obscure the user's view of the external environment, the mirror 210 can be made

partially reflective so that the user sees the external environment, through the mirror 210, with the virtual image superposed over the real external environment.

In the case where separate respective displays are provided for each of the user's eyes, it is possible to display stereoscopic images. An example of a pair of stereoscopic images for display to the left and right eyes is shown in Figure 5. The images exhibit a lateral displacement relative to one another, with the displacement of image features depending upon the (real or simulated) lateral separation of the cameras by which the images were captured, the angular convergence of the cameras and the (real or simulated) distance of each image feature from the camera position.

Note that the lateral displacements in Figure 5 could in fact be the other way round, which is to say that the left eye image as drawn could in fact be the right eye image, and the right eye image as drawn could in fact be the left eye image. This is because some stereoscopic displays tend to shift objects to the right in the right eye image and to the left in the left eye image, so as to simulate the idea that the user is looking through a stereoscopic window onto the scene beyond. However, some HMDs use the arrangement shown in Figure 5 because this gives the impression to the user that the user is viewing the scene through a pair of binoculars. The choice between these two arrangements is at the discretion of the system designer.

In some situations, an HMD may be used simply to view movies and the like. In this case, there is no change required to the apparent viewpoint of the displayed images as the user turns the user's head, for example from side to side. In other uses, however, such as those associated with virtual reality (VR) or augmented reality (AR) systems, the user's viewpoint needs to track movements with respect to a real or virtual space in which the user is located.

As mentioned above, in some uses of the HMD, such as those associated with virtual reality (VR) or augmented reality (AR) systems, the user's viewpoint needs to track movements with respect to a real or virtual space in which the user is located.

This tracking is carried out by detecting motion of the HMD and varying the apparent viewpoint of the displayed images so that the apparent viewpoint tracks the motion. The detection may be performed using any suitable arrangement (or a combination of such arrangements). Examples include the use of hardware motion detectors (such as accelerometers or gyroscopes), external cameras operable to image the HMD, and outwards-facing cameras mounted onto the HMD.

Turning to gaze tracking in such an arrangement, Figure 6 schematically illustrates two possible arrangements for performing eye tracking on an HMD. The cameras provided within such arrangements may be selected freely so as to be able to perform an effective eye-tracking method. In some existing arrangements, visible light cameras are used to capture images of a user's eyes. Alternatively, infra-red (IR) cameras are used so as to reduce interference either in

the captured signals or with the user's vision should a corresponding light source be provided, or to improve performance in low-light conditions.

Figure 6a shows an example of a gaze tracking arrangement in which the cameras are arranged within an HMD so as to capture images of the user's eyes from a short distance. This may be referred to as near-eye tracking, or head-mounted tracking.

In this example, an HMD 600 (with a display element 601) is provided with cameras 610 that are each arranged so as to directly capture one or more images of a respective one of the user's eyes using an optical path that does not include the lens 620. This may be advantageous in that distortion in the captured image due to the optical effect of the lens is able to be avoided. Four cameras 610 are shown here as examples of possible positions that eye-tracking cameras may provided, although it should be considered that any number of cameras may be provided in any suitable location so as to be able to image the corresponding eye effectively. For example, only one camera may be provided per eye or more than two cameras may be provided for each eye.

However it is considered that in a number of embodiments it is advantageous that the cameras are instead arranged so as to include the lens 620 in the optical path used to capture images of the eye. Examples of such positions are shown by the cameras 630. While this may result in processing being required to enable suitably accurate tracking to be performed, due to the deformation in the captured image due to the lens, this may be performed relatively simply due to the fixed relative positions of the corresponding cameras and lenses. An advantage of including the lens within the optical path may be that of simplifying the physical constraints upon the design of an HMD, for example.

Figure 6b shows an example of a gaze tracking arrangement in which the cameras are instead arranged so as to indirectly capture images of the user's eyes. Such an arrangement may be particularly suited to use with IR or otherwise non-visible light sources, as will be apparent from the below description.

Figure 6b includes a mirror 650 arranged between a display 601 and the viewer's eye (of course, this can be extended to or duplicated at the user's other eye as appropriate). For the sake of clarity, any additional optics (such as lenses) are omitted in this Figure – it should be appreciated that they may be present at any suitable position within the depicted arrangement. The mirror 650 in such an arrangement is selected so as to be partially transmissive; that is, the mirror 650 should be selected so as to enable the camera 640 to obtain an image of the user's eye while the user views the display 601. One method of achieving this is to provide a mirror 650 that is reflective to IR wavelengths but transmissive to visible light – this enables IR light used for tracking to be reflected from the user's eye towards the camera 640 while the light emitted by the display 601 passes through the mirror uninterrupted.

Such an arrangement may be advantageous in that the cameras may be more easily arranged out of view of the user, for instance. Further to this, improvements to the accuracy of the eye tracking may be obtained due to the fact that the camera captures images from a position that is effectively (due to the reflection) along the axis between the user's eye and the display.

Of course, eye-tracking arrangements need not be implemented in a head-mounted or otherwise near-eye fashion as has been described above. For example, Figure 7 schematically illustrates a system in which a camera is arranged to capture images of the user from a distance; this distance may vary during tracking, and may take any value in dependence upon the parameters of the tracking system. For example, this distance may be thirty centimetres, a metre, five metres, ten metres, or indeed any value so long as the tracking is not performed using an arrangement that is affixed to the user's head.

In Figure 7, an array of cameras 700 is provided that together provide multiple views of the user 710. These cameras are configured to capture information identifying at least the direction in which a user's 710 eyes are focused, using any suitable method. For example, IR cameras may be utilised to identify reflections from the user's 710 eyes. An array of cameras 700 may be provided so as to provide multiple views of the user's 710 eyes at any given time, or may be provided so as to simply ensure that at any given time at least one camera 700 is able to view the user's 710 eyes. It is apparent that in some use cases it may not be necessary to provide such a high level of coverage and instead only one or two cameras 700 may be used to cover a smaller range of possible viewing directions of the user 710.

Of course, the technical difficulties associated with such a long-distance tracking method may be increased; higher resolution cameras may be required, as may stronger light sources for generating IR light, and further information (such as head orientation of the user) may need to be input to determine a focus of the user's gaze. The specifics of the arrangement may be determined in dependence upon a required level of robustness, accuracy, size, and/or cost, for example, or any other design consideration.

Despite technical challenges including those discussed above, such tracking methods may be considered beneficial in that they allow a greater range of interactions for a user – rather than being limited to HMD viewing, gaze tracking may be performed for a viewer of a television, for instance.

Rather than varying only in the location in which cameras are provided, eye-tracking arrangements may also differ in where the processing of the captured image data to determine tracking data is performed.

Figure 8 schematically illustrates an environment in which an eye-tracking process may be performed. In this example, the user 800 is using an HMD 810 that is associated with the processing unit 830, such as a games console, with the peripheral 820 allowing a user 800 to

input commands to control the processing. The HMD 810 may perform eye tracking in line with an arrangement exemplified by Figure 6a or 6b, for example – that is, the HMD 810 may comprise one or more cameras operable to capture images of either or both of the user's 800 eyes. The processing unit 830 may be operable to generate content for display at the HMD 810; although some (or all) of the content generation may be performed by processing units within the HMD 810.

The arrangement in Figure 8 also comprises a camera 840, located outside of the HMD 810, and a display 850. In some cases, the camera 840 may be used for performing tracking of the user 800 while using the HMD 810, for example to identify body motion or a head orientation. The camera 840 and display 850 may be provided as well as or instead of the HMD 810; for example these may be used to capture images of a second user and to display images to that user while the first user 800 uses the HMD 810, or the first user 800 may be tracked and view content with these elements instead of the HMD 810. That is to say, the display 850 may be operable to display generated content provided by the processing unit 830 and the camera 840 may be operable to capture images of one or more users' eyes to enable eye-tracking to be performed.

While the connections shown in Figure 8 are shown by lines, this should of course not be taken to mean that the connections should be wired; any suitable connection method, including wireless connections such as wireless networks or Bluetooth®, may be considered suitable. Similarly, while a dedicated processing unit 830 is shown in Figure 8 it is also considered that the processing may in some embodiments be performed in a distributed manner – such as using a combination of two or more of the HMD 810, one or more processing units, remote servers (cloud processing), or games consoles.

The processing required to generate tracking information from captured images of the user's 800 eye or eyes may be performed locally by the HMD 810, or the captured images or results of one or more detections may be transmitted to an external device (such as a the processing unit 830) for processing. In the former case, the HMD 810 may output the results of the processing to an external device for use in an image generation process if such processing is not performed exclusively at the HMD 810. In embodiments in which the HMD 810 is not present, captured images from the camera 840 are output to the processing unit 830 for processing.

Figure 9 schematically illustrates a system for performing one or more eye tracking processes, for example in an embodiment such as that discussed above with reference to Figure 8. The system 900 comprises a processing device 910, one or more peripherals 920, an HMD 930, a camera 940, and a display 950. Of course, not all elements need be present within the system 900 in a number of embodiments – for instance, if the HMD 930 is present then it is

considered that the camera 940 may be omitted as it is unlikely to be able to capture images of the user's eyes.

As shown in Figure 9, the processing device 910 may comprise one or more of a central processing unit (CPU) 911, a graphics processing unit (GPU) 912, storage (such as a hard drive, or any other suitable data storage medium) 913, and an input/output 914. These units may be provided in the form of a personal computer, a games console, or any other suitable processing device.

For example, the CPU 911 may be configured to generate tracking data from one or more input images of the user's eyes from one or more cameras, or from data that is indicative of a user's eye direction. This may be data that is obtained from processing images of the user's eye at a remote device, for example. Of course, should the tracking data be generated elsewhere then such processing would not be necessary at the processing device 910.

The GPU 912 may be configured to generate content for display to the user on which the eye tracking is being performed. In some embodiments, the content itself may be modified in dependence upon the tracking data that is obtained – an example of this is the generation of content in accordance with a foveal rendering technique. Of course, such content generation processes may be performed elsewhere – for example, an HMD 930 may have an on-board GPU that is operable to generate content in dependence upon the eye tracking data.

The storage 913 may be provided so as to store any suitable information. Examples of such information include program data, content generation data, and eye tracking model data. In some cases, such information may be stored remotely such as on a server, and as such a local storage 913 may not be required – the discussion of the storage 913 should therefore be considered to refer to local (and in some cases removable storage media) or remote storage.

The input/output 914 may be configured to perform any suitable communication as appropriate for the processing device 910. Examples of such communication include the transmission of content to the HMD 930 and/or display 950, the reception of eye-tracking data and/or images from the HMD 930 and/or the camera 940, and communication with one or more remote servers (for example, via the internet).

As discussed above, the peripherals 920 may be provided to allow a user to provide inputs to the processing device 910 in order to control processing or otherwise interact with generated content. This may be in the form of button presses or the like, or alternatively via tracked motion to enable gestures to be used as inputs.

The HMD 930 may comprise a number of sub-elements, which have been omitted from Figure 9 for the sake of clarity. Of course, the HMD 930 should comprise a display unit operable to display images to a user. In addition to this, the HMD 930 may comprise any number of suitable cameras for eye tracking (as discussed above), in addition to one or more processing

units that are operable to generate content for display and/or generate eye tracking data from the captured images.

The camera 940 and display 950 may be configured in accordance with the discussion of the corresponding elements above with respect to Figure 8.

Turning to the image capture process upon which the eye tracking is based, examples of different cameras are discussed. The first of these is a standard camera, which captures a sequence of images of the eye that may be processed to determine tracking information. The second is that of an event camera, which instead generates outputs in response to observed changes in the incident light, as discussed later.

Traditional image-based gaze tracking techniques use standard cameras given that they are widely available and often relatively cheap to produce. 'Standard cameras' here refer to cameras which capture images of the environment at predetermined intervals which can be combined to generate video content. For example, a typical camera of this type may capture thirty image frames each second, and these images may be output to a processing unit for feature analysis or the like to be performed so as to enable tracking of the eye.

Such a camera comprises a light-sensitive array that is operable to record light information during an exposure time, with the exposure time being controlled by a shutter speed (the speed of which dictates the frequency of image capture). The shutter may be configured as a rolling shutter (line-by-line reading of the captured information) or a global shutter (reading the captured information of the whole frame simultaneously), for example.

Independent of the type of camera that is selected, in many cases it may be advantageous to provide illumination to the eye in order to obtain a suitable image. One example of this is the provision of an IR light source that is configured to emit light in the direction of one or both of the user's eyes; an IR camera may then be provided that is able to detect reflections from the user's eye in order to generate an image. IR light may be preferable as it is invisible to the human eye, and as such does not interfere with normal viewing of content by the user, but it is not considered to be essential. In some cases, the illumination may be provided by a light source that is affixed to the imaging device, while in other embodiments it may instead be that the light source is arranged away from the imaging device.

As suggested in the discussion above, the human eye does not have a uniform structure; that is, the eye is not a perfect sphere, and different parts of the eye have different characteristics (such as varying reflectance or colour). Figure 10 shows a simplified side view of the structure of a typical eye 1000; this Figure has omitted features such as the muscles which control eye motion for the sake of clarity.

The eye 1000 is formed of a near-spherical structure filled with an aqueous solution 1010, with a retina 1020 formed on the rear surface of the eye 1000. The optic nerve 1030 is connected at the rear of the eye 1000. Images are formed on the retina 1020 by light entering

the eye 1000, and corresponding signals carrying visual information are transmitted from the retina 1020 to the brain via the optic nerve 1030.

Turning to the front surface of the eye 1000, the sclera 1040 (commonly referred to as the white of the eye) surrounds the iris 1050. The iris 1050 controls the size of the pupil 1060, which is an aperture through which light enters the eye 1000. The iris 1050 and pupil 1060 are covered by the cornea 1070, which is a transparent layer which can refract light entering the eye 1000. The eye 1000 also comprises a lens (not shown) that is present behind the iris 1050 that may be controlled to adjust the focus of the light entering the eye 1000.

The structure of the eye is such that there is an area of high visual acuity (the fovea), with a sharp drop off either side of this. This is illustrated by the curve 1100 of Figure 11, with the peak in the centre representing the foveal region. The area 1110 is the 'blind spot'; this is an area in which the eye has no visual acuity as it corresponds to the area where the optic nerve meets the retina. The periphery (that is, the viewing angles furthest from the fovea) is not particularly sensitive colour or detail, and instead is used to detect motion.

As has been discussed above, foveal rendering is a rendering technique that takes advantage of the relatively small size (around 2.5 degrees) of the fovea and the sharp fall-off in acuity outside of that.

The eye undergoes a large amount of motion during viewing, and this motion may be categorised into one of a number of categories.

A saccadic eye movement is identified as a fast motion of the eye in which the eye moves in a ballistic manner to abruptly change a point of fixation. This may be considered as ballistic movement, in that once the movement of the eye has been initiated to change a point of focus from a current point of focus to a target point of focus (next point of focus), the target point of focus and the direction of movement of the eye to move the point of focus to the target point of focus cannot be altered by the human visual system. As such, during the course of the eye movement to change the saccade from the current fixation point to the next fixation point for the eye it is not possible to interrupt the eye movement, and upon reaching the target fixation point the eye remains stationary for a period of time (a fixation pause) to focus on the target fixation point before subsequent eye movement can be initiated. It is sometimes observed that a saccade is followed by a second smaller corrective saccade that is performed to bring the eye closer to the target fixation point. Such a corrective saccade typically occurs after a very short period of time. A saccade can range in size from a small eye movement made while reading, for example, to a much larger eye movement made when observing a surrounding environment. Saccades are often not conscious eye movements, and instead are performed reflexively to focus on a target when surveying an environment. Saccades may last up to two hundred milliseconds, depending on the angle rotated by the eye to change the position of the fovea and thus the foveal region of the viewer's vision to thereby change the point of fixation for the eye,

but may be as short as twenty milliseconds. The rotational speed of the eye during a saccade is also dependent upon a magnitude of a total rotation angle of the eye; typical speeds may range from two hundred to five hundred degrees per second.

'Smooth pursuit' refers to a slower movement type than a saccade. Smooth pursuit is generally associated with a conscious tracking of a point of focus by a viewer, and is performed so as to maintain the position of a target within (or at least substantially within) the foveal region of the viewer's vision. This enables a high-quality view of a target of interest to be maintained in spite of motion. If the target moves too fast, then smooth pursuit may instead require a number of saccades in order to keep up; this is because smooth pursuit has a lower maximum speed, in the region of thirty degrees per second.

The vestibular-ocular reflex is a further example of eye motion. The vestibular-ocular reflex is the motion of the eyes that counteracts head motion; that is, the motion of the eyes relative to the head that enables a person to remain focused on a particular point despite moving their head.

Another type of motion is that of the vergence accommodation reflex. This is the motion that causes the eyes to rotate to converge at a point, and the corresponding adjustment of the lens within the eye to cause that point to come into focus.

Further eye motions that may be observed as a part of a gaze tracking process are those of blinks or winks, in which the eyelid covers the eyes of the user.

Movements of the eye are performed by a user wearing an HMD whilst viewing images displayed by the HMD to enable detailed visual analysis of a portion of an image displayed by the HMD. In particular, the eye can be rotated to reposition the fovea and the pupil to enable detailed visual analysis for the portion of the image for which light is incident upon the fovea. Similarly, movements of the eye are also performed by a user not wearing an HMD whilst viewing images displayed by a display unit, such as the display unit 850 or 950 described previously with reference to Figures 8 and 9.

Conventional techniques for foveated rendering typically require multiple render passes to allow an image frame to be rendered multiple times at different image resolutions so that the resulting renders are then composited together to achieve regions of different image resolution in an image frame. The use of multiple render passes requires significant processing overhead and undesirable image artefacts can arise at the boundaries between the regions. Alternatively, in some cases hardware can be used that allows rendering at different resolutions in different parts of an image frame without needing additional render passes. Such hardware-accelerated implementations may therefore be better in terms of performance, but this comes with limitations as to the smoothness of the transition between the regions of different image resolution within the image frame. In some implementations, only a limited number of regions can be used and a noticeably sharp drop in image resolution is observed between the regions.

The operations to be discussed below relate to using machine learning to enhance an image frame by predicting pixel values for the image frame. In particular, first pixel values included in the image frame are obtained and then input to a machine learning model, and a gaze point detected for a user's eye is also input to the machine learning model. The machine learning model is trained to predict second pixel values for the image frame for enhancing the image frame and a predicted image frame can be stored comprising the predicted second pixel values and some of the first pixel values. Therefore, a predicted image frame can be obtained which comprises at least a portion that has a higher image resolution than the image resolution of that portion in the image frame. In addition to enhancing the image resolution for a portion of the input image frame, other possibilities for further enhancing the input image frame are discussed below.

Figure 12 schematically illustrates a data processing system 1200 for predicting second pixel values and storing a predicted image frame in dependence upon first pixel values in an input image frame and a gaze point of a user for the input image frame. In embodiments of the disclosure, the data processing system 1200 comprises: processing circuitry 1210 to obtain first pixel values for an image frame; input circuitry 1220 to receive an input indicative of a gaze point of an eye of a user for the image frame; a machine learning model 1230 trained to predict second pixel values for a portion of the image frame in dependence upon the first pixel values and the gaze point, the machine learning model 1230 trained with training image frames each comprising a first portion having a first image resolution and a second portion having a second image resolution, the first image resolution being higher than the second image resolution, in which a number of the respective second pixel values predicted for the portion of the image frame is greater than a number of the respective first pixel values obtained by the processing circuitry for the portion of the image frame; storage circuitry 1240 to store a predicted image frame comprising the second pixel values and some of the first pixel values obtained by the processing circuitry for the image frame, in which the second pixel values correspond to a high image resolution portion of the predicted image frame; and output circuitry 1250 to output the predicted image frame for display to the user. The data processing system 1200 may be provided as part of a processing device, such as the processing device 910, or provided as part of an HMD 600 810 or as part of a server.

In the case where the data system 1200 is provided as part of the processing device 910, the input circuitry 1220 can receive the input comprising information indicative of the gaze point of the eye of the user for the image frame via a wired or wireless communication (e.g. Bluetooth® communication link) from an HMD comprising a gaze detector (such as the HMD 600, 810) or from a detector (such as any one of the detectors 610, 630, 640, 700, 840, 940) and the output circuitry 1250 can output the predicted image frame for display to the user by communicating image data corresponding to the predicted image frame to the HMD or a display

unit (such as the display unit 950) arranged with respect to the user via a wired or wireless communication. In some examples, the data processing system 1200 may be provided as part of a server, the input circuitry 1220 can be configured to receive an input comprising information indicative of the gaze point of the eye of the user for the image frame from the HMD or the detector (or a processing device, such as a personal computer or a game console associated with the HMD or the detector) via a wireless communication, and the output circuitry 1250 can be configured to output the predicted image frame for display to the user by communicating image data corresponding to the predicted image frame to the HMD or a display unit (such as the display unit 950) arranged with respect to the user.

The processing circuitry 1210 is configured to obtain the first pixel values for the image frame. In some embodiments of the disclosure, the processing circuitry 1210 is configured to obtain first pixel values for each of a plurality of successive image frames, as discussed in more detail later. Each image frame comprises a given number of image pixels arranged in rows and columns. For example, a 1080p image frame comprises 1920 x 1080 image pixels. Each image pixel has at least one corresponding pixel value. For a grayscale image, each pixel may have a single 4-bit, 8-bit or 16-bit data value, for example, indicative of an intensity for the image pixel. For colour images, each image pixel may have one or more corresponding N-bit data values (e.g. 8-bit data values) for the respective colour channels (e.g. RGB, YCbCr) for that image pixel (where N is an integer greater than or equal to two). In some examples, each image pixel may have a 24-bit number including an 8-bit number for each of the red, green and blue colour components. In other examples, the image frame may have a 30-bit RGB format in which each colour channel has a 10-bit data value. It will be appreciated that other image formats may be used and the type of image format is not limited.

The processing circuitry 1210 is configured to obtain the first pixel values for an image frame either by decoding the image frame or by performing rendering operations as part of an execution of an application such as a computer game to render the image frame. In embodiments of the disclosure, the processing circuitry 1210 comprises either rendering circuitry to perform rendering operations to obtain the first pixel values or decoding circuitry to perform decoding operations to obtain the first pixel values from compressed image data. Rendering operations typically comprise processing of model data or other predefined graphical data to obtain pixel values for the image pixels in the image frame. The rendering circuitry (which can be provided as part of a GPU and/or a CPU) is configured to perform rendering operations for the image frame to obtain the first pixel values for the image frame. The rendering circuitry can be configured to render respective image frames such that the first pixel values can be obtained for each of the image frames and the first pixel values for the image frame can be input to the machine learning model 1220. Alternatively, the processing circuitry 1210 may comprise decoding circuitry (which can be provided as part of a GPU and/or a CPU)

to perform decoding operations for compressed image data to obtain the first pixel values. For example, the processing circuitry 1210 can be configured to receive the image data via a wired or wireless communication (e.g. from a remote server) and the decoding circuitry can be configured to decode the image data to obtain pixel values for each of the image pixels in an image frame. In some examples, the data processing system 1200 may comprise a memory (not shown in Figure 12) configured to store image data for a plurality of image frames. Various types of video decoding are considered and the type of decoding is not limited.

Hence more generally, the processing circuitry 1210 is configured to obtain a number of respective first pixel values for a given image frame, and the respective first pixel values are provided as an input to the trained machine learning model 1230. References herein to images frames refer to either stereoscopic image frames comprising left and right images, or a single image frame that is to be viewed by both eyes of the user.

The input circuitry 1220 is configured to receive the input indicative of the gaze point of the eye of the user for the image frame. A detector can be arranged with respect to the user to detect a gaze point for the user. Alternatively, when the user is wearing an HMD, one or more detectors proved as part of the HMD can detect the gaze point for the user. Information indicative of the gaze point for the user can be communicated to the input circuitry 1220 via a wired or wireless communication from at least one of the HMD 600, 810 and any one of the detectors 610, 630, 640, 700, 840, 940. Examples of suitable wireless connections include a Bluetooth® connection. Therefore, the input circuitry 1220 can receive an input indicative of the gaze point and information indicative of the gaze point can be provided as an input to the trained machine learning model 1230. In some cases the data processing system 1200 may be provided as part of a server and the data processing system 1200 may receive the input indicative of the gaze point from a personal computer or a game console associated with the HMD or the detector.

In the data processing system of Figure 12, the processing circuitry 1210 obtains first pixel values for the image frame, the input circuitry 1220 receives the input indicative of the gaze point and, on the basis of the first pixel values and the gaze point for the image frame, the machine learning model 1230 predicts the second pixel values for a portion of the image frame so as to enhance an image resolution for the portion of the image frame. In this way, a predicted image frame is obtained which differs from the input image frame in that a number of pixel values corresponding to the portion is greater for the predicted image frame than for the input image frame. Therefore, the data processing system 1200 can allow the input image frame (initial image frame) to be upscaled in a specific region by predicting the second pixel values for the specific region, where the location of the specific region is identified in dependence upon the gaze point received by the input circuitry 1220 for the input image frame. As such, rather than upscaling the entire image frame, the data processing system 1200 can predict the second pixel

values so as to allow upscaling of the region of the image frame corresponding to the user's gaze point thereby improving processing efficiency. Therefore, the image can be effectively increased in image resolution in the portion for which the user has high visual acuity. Hence, a number of the respective second pixel values predicted for the portion of the image frame is greater than a number of the respective first pixel values obtained by the processing circuitry 1210 for that same portion of the image frame and therefore at least that portion of the image frame is enhanced.

The location of the portion of the image frame for which the second pixel values are predicted is dependent upon the gaze point. The size and shape of the portion is dependent upon the properties of the training image frames which are discussed in more detail later.

The storage circuitry 1230 is configured to store the predicted image frame, in which the predicted image frame comprises the second pixel values predicted by the machine learning model 1230 and at least some of the first pixel values obtained by the processing circuitry 1210 for the initial image frame. The second pixel values in the predicted image frame correspond to a high image resolution portion of the predicted image frame. The remaining part of the predicted image frame is of lower image resolution. As discussed in more detail later, the part of the predicted image frame outside the high image resolution portion corresponding to the second pixel values may comprise just the first pixel values obtained by the processing circuitry 1210 for that part of the image frame, or in some cases third pixel values may also be predicted for another portion of the predicted image frame that at least partially surrounds the high image resolution portion corresponding to the second pixel values. As such, the first pixel values obtained by the processing circuitry 1210 for the high image resolution portion of the predicted image frame can be replaced with the predicted second pixel values, and for the remaining part of the predicted image frame the first pixel values obtained by the processing circuitry 1210 can be retained. The storage circuitry 1230 comprises one or more from the list consisting of: a buffer memory, a cache memory, a framebuffer memory and a flash memory.

Figure 13a schematically illustrates an example of a predicted image frame 1300 comprising a high image resolution portion 1310 including the second pixel values and a remaining portion 1320 including some of the first pixel values obtained by the processing circuitry 1210. The initial image frame for which the first pixel values are obtained has a first image resolution (e.g. a 720p image frame or possibly an image frame having a much lower image resolution, or a sparse distribution of pixels in the form of e.g. a 50% chequerboard or 1 in 4 pixels within a 1080p image), and the machine learning model 1230 can be trained to predict the second pixel values for the initial image frame such that a number of the respective second pixel values predicted for the portion 1310 is such that the number of respective second pixel values per unit area in the portion 1310 is greater than the number of respective first pixel values per unit area in the initial image frame. In other words, the number of the respective

second pixel values predicted for the portion 1310 is such that the pixel value density (number of pixels per unit area) is greater in the portion 1310 in the predicted image frame 1300 than for that same portion in the initial lower resolution image frame. The portion 1310 of the predicted image frame 1300 is positioned so that the gaze point is located at the centre of the portion 1310. Consequently, the portion 1310 has a higher image resolution and the remaining portion 1320 has a lower image resolution that is the same as the image resolution of the initial image frame.

The size and the shape of the portion 1310 including the second pixel values is not limited to that shown in Figure 13a. The size and shape of the portion 1310 is dependent upon the properties of the training image frame used for training the machine learning model 1230. The portion 1310 may take a variety of shapes such as a circle, oval, square or rectangle for example. The training image frames can be generated by post-processing image frames generated for a given video content (e.g. a video game or a film or recorded television programme) so that a high resolution portion in the training image frame can take any suitable shape when creating the training images, and the geometric size of the high resolution portion in the training image frame can also take any suitable size. The high resolution portion is smaller in size than the training image frame and thus represents a portion of the training image frame. For example, the size of the high resolution portion in the training image frame can be selected so as to correspond to at least the portion of the image observed by the fovea of the user's eye. As has been discussed above, foveal rendering is a rendering technique that takes advantage of the relatively small size (around 2.5 degrees) of the fovea and the sharp fall-off in resolution acuity outside of that. Therefore, the high resolution portion of the training image frame can be selected so as to correspond to at least 5 degrees of the visual field for the eye. Further details regarding the creation of the training image frames are discussed later.

Referring again to Figure 12, the output circuitry 1250 is configured to output the predicted image frame for display to the user. The output circuitry 1250 can be configured to access data stored by the storage circuitry 1240 and to output the predicted image frame to an HMD or a display unit (such as 950) or another processing device (e.g. processing device 910) for display to the user. The output circuitry 1250 is configured to output the predicted image frame for display via a wired or wireless communication.

In embodiments of the disclosure, the processing circuitry 1210 is configured to obtain the first pixel values for each of a plurality of image frames and the output circuitry 1240 is configured to output a predicted image frame for each of the image frames. The processing circuitry 1210 can receive an encoded stream of image data and decode the stream of image data to obtain the first pixel values for each image frame in a sequence of image frames. Alternatively, the processing circuitry 1210 can perform rendering operations in accordance with a computer graphics pipeline to obtain the first pixel values for each image frame in a sequence

of image frames. The first pixel values can thus be obtained for each image frame and provided as an input to the machine learning model 1230 along with the gaze point detected for that image frame. In this way, first pixel values can be acquired for each of a plurality of successive low resolution images frames and the output circuitry 1250 can output a plurality of successive predicted image frames each comprising a high resolution portion corresponding to the gaze point and including the second pixel values.

An input comprising information indicative of the gaze point is received by the input circuitry 1220. In some examples, the data processing system 1200 further comprises a detector (such as any one of the detectors 610, 630, 640, 700, 840, 940) arranged with respect to the user to detect the gaze point for the user and communicate the information indicative of the gaze point to the input circuitry 1220. For example, the detector may be a camera that captures images of the user's eye according to a fixed frame rate or may be an event camera, as discussed previously. In some examples, the data processing system 1200 further comprises an HMD for detecting the user's gaze point. The information indicative of the gaze point for the user is thus received by the input circuitry 1220. Upon completion of the processing by the processing circuitry 1210 to obtain the first pixel values for a given image frame, the first pixel values and a most recently received gaze point are provided to the machine learning model 1230. Therefore, in some examples, the input circuitry 1220 can receive the information indicative of the user's gaze point during the processing by the processing circuitry 1210 for obtaining the first pixel values for the given image frame. Consequently, information indicative of the user's gaze direction can be received by the input circuitry 1220 and a gaze point detected with respect to one or more previous image frames output for display to the user can be provided as an input to the machine learning model 1230 together with the first pixel values for a current image frame for which second pixel values are to be predicted.

In embodiments of the disclosure, the machine learning model 1230 is trained to predict third pixel values for another portion of the image frame in dependence upon the first pixel values and the gaze point, in which the another portion of the image frame at least partially surrounds the portion of the image frame for which the second pixel values are predicted, the training image frames each comprising a third portion between the first portion and the second portion, the third portion corresponding to a transition region and having an image resolution higher than the second image resolution and lower than the first image resolution. In addition to predicting the second pixel values for the portion of the image frame centred upon the gaze direction, the machine learning model 1230 can be trained to predict third pixel values for another portion of the image frame, where the another portion of the image frame is arranged so as to either completely or partially surround the portion for which the second pixel values are predicted.

Figure 13b schematically illustrates an example of a predicted image frame 1350 comprising a high image resolution portion 1310 including the second pixel values, another portion 1360 (also referred to as an intermediate image resolution portion) including the third pixel values, and a remaining portion 1370 including some of the first pixel values obtained by the processing circuitry 1210 for the input image frame corresponding to the predicted image frame 1350. Therefore, the predicted image frame comprises the three portions 1310, 1360 1370 where the portion 1360 includes the third pixel values such that the portion 1360 represents an intermediate image resolution portion 1360 for transitioning between the high image resolution portion 1310 and the low image resolution portion 1370. This can avoid the appearance of an abrupt change in image resolution which may otherwise occur if the high image resolution portion 1310 and the low image resolution portion 1370 share a boundary. By predicting the third pixel values, an image resolution can be achieved for the intermediate image resolution portion 1360 such that the image resolution does not transition directly from the high image resolution to the low image resolution, but rather a transition region is provided having an intermediate image resolution. In addition, by providing such an intermediate image resolution portion 1360, it may be possible to use a smaller size for the high resolution portion 1310 in Figure 13b compared to the case shown in Figure 13a, because a combination of the high resolution portion 1310 and the intermediated resolution portion 1360 can be used instead to achieve the same coverage whilst ensuring that the high image resolution portion 1310 is centred upon the gaze point.

In some examples, the intermediate image resolution portion 1360 has a uniform image resolution such that a pixel value density is substantially the same throughout the portion 1360. For example, the intermediate image resolution portion 1360 may have an image resolution that represents a midpoint image resolution between the image resolution of the high image resolution portion 1310 and the image resolution of the low image resolution portion 1370. For example, in the case where the high image resolution portion 1310 has an image resolution R1 and the low image resolution portion 1370 has an image resolution R2, the image resolution of the intermediate image resolution portion may be given by: $R3 = (R1+R2)/2$.

Alternatively, the intermediate image resolution portion 1360 may have an image resolution that is more similar to the image resolution of the high image resolution portion 1310 than the image resolution of the low image resolution portion 1370. In other words, the high image resolution portion 1310 may have a first image resolution, the intermediate image resolution portion 1360 may have a second image resolution and the low image resolution portion 1370 may have a third image resolution, in which a difference between the first image resolution and the second image resolution is smaller than a difference between the second image resolution and the third image resolution. The change in human visual resolution acuity as a function of distance with respect to the gaze point follows the trend as shown in Figure 11,

and there is a sharp fall off in resolution acuity outside the fovea. Consequently, the high image resolution portion 1310 can be defined so as to cover at least 5 degrees of the visual field for the eye centred upon the gaze point and the intermediate image resolution portion 1360 can, for example, be defined so as to at least partially surround the high image resolution portion 1310 and cover approximately 15 degrees of the visual field for the eye with an image resolution that is greater than a midpoint image resolution representing a midpoint between the high image resolution and the low image resolution.

Figure 13b illustrates an example in which the gaze point of the user corresponds to the centre of the input image frame and thus the centre of the predicted image frame 1350, such that the portion 1310 is centred upon the gaze point. In the case shown in Figure 13b, the another portion 1360 completely surrounds the portion 1310 for which the second pixel values are predicted. However, the user's gaze point may be positioned differently to that shown in Figure 13b. For the case where the gaze point is offset with respect to the centre of the input image frame and positioned proximate to a periphery of the input image frame, the another portion 1360 may only partially surround the portion 1310.

In embodiments of the disclosure, a number of the respective third pixel values predicted for the another portion of the image frame is greater than a number of the respective first pixel values obtained by the processing circuitry 1210 for the another portion of the image frame. The machine learning model 1230 is trained to predict the third pixel values for the another portion of the input image frame in dependence upon the first pixel values and the gaze point, such that a number of the respective third pixel values exceeds a number of the respective first pixel values obtained by the processing circuitry 1210 for the another portion of the input image frame. Therefore, a number of the respective third pixel values in the intermediate image resolution portion 1360 in the predicted image frame 1350 exceeds a number of the respective first pixel values obtained by the processing circuitry 1210 for the another portion of the input image frame. The another portion of the input image frame corresponds to the intermediate image resolution portion 1360 in the predicted image frame 1350 in that they have the same shape and size. Therefore, by predicting the third pixel values, can be increased for the another portion of the input image frame. The machine learning model 1230 can be trained to predict the second pixel values and the third pixel values such that a number of the respective second pixel values predicted for the portion 1310 is such that the number of respective second pixel values per unit area in the portion 1310 is greater than the number of respective first pixel values per unit area in the initial image frame, and a number of the respective third pixel values predicted for the portion 1360 is such that the number of respective third pixel values per unit area in the portion 1360 is greater than the number of respective first pixel values per unit area in the initial image frame and is less than (or at least does not exceed) the number of respective second pixel values per unit area in the portion 1310.

In embodiments of the disclosure, the second pixel values in the predicted image frame 1350 correspond to a high image resolution portion of the predicted image frame, the third pixel values in the predicted image frame 1350 correspond to an intermediate image resolution portion of the predicted image frame, and the first pixel values in the predicted image frame correspond to a low image resolution portion of the predicted image frame. For the predicted image frame 1350, the first pixel values obtained by the processing circuitry 1210 for the high image resolution portion 1310 of the predicted image frame 1350 can be replaced with (substituted for) the second pixel values predicted by the machine learning model 1230, the first pixel values obtained by the processing circuitry 1210 for the intermediate image resolution portion 1360 of the predicted image frame 1350 can be replaced with the third pixel values predicted by the machine learning model 1230, and for the remaining portion 1370 of the predicted image frame 1250 the first pixel values obtained by the processing circuitry 1210 can be retained. Therefore, the storage circuitry 1240 stores the predicted image frame 1350 comprising the second pixel values, the third pixel values and the first pixel values corresponding to the low image resolution portion.

Whilst Figure 13b shows the portion 1310 and the another portion 1360, it will be appreciated that the relative sizes of the two portions 1310, 1360 is not limited to that shown and is dependent upon the properties of the training image frames. The training image frames each comprise a third portion between the first portion and the second portion, where the third portion has an image resolution lower than the image resolution of the first portion and higher than the image resolution of the second portion such that the third portion represents a transition region in the training image frame. The image resolution of the third portion in the training image frame may be uniform or may vary with respect to distance from the centre point of the first portion such that the third portion has a varying image resolution. In the case of varying image resolution, the third portion may have an image resolution less than or equal to the image resolution of the first portion (high resolution portion) at a given distance from the centre point of the first portion and may also have an image resolution higher than or equal to the image resolution of the second portion (low resolution portion) at another given distance from the centre point such that the third portion represents a transition region in the training image frame between the first portion and the second portion. This is discussed in more detail later.

As explained previously with respect to Figure 13a, the training image frames can be generated by post-processing image frames generated for a given video content so as to allow a size (and therefore a relative size of the first, second and third portions) of the third portion to be selected. Therefore, the high resolution portion (first portion) and the intermediate resolution portion (third portion) in the training image frames can take any suitable shape and can also take any suitable size within the training image frame. Typically, when a first type of shape is

used for the first portion in the training image frame a similar type of shape is used for the third portion in the training image frame. For example, as shown in Figure 13b, the portion 1310 has circular shape with the portion 1360 having ring shape with a size that allows the portion 1360 to at least partially surround the portion 1310. However, in some examples, a first type of shape may be used for the high image resolution portion 1310, and a second type of shape different from the first type of shape may be used for the intermediate image resolution portion 1360. For example, a circular shape may be used for the portion 1310 whereas a square shape may be used for the portion 1360 to at least partially surround the portion 1310.

In embodiments of the disclosure, the image resolution of the intermediate image resolution portion 1360 in the predicted image frame 1350 varies in dependence upon a distance from the gaze point, and the image resolution of the third portion of each training image frame varies in dependence upon a distance from a centre point of the first portion. Whereas the high image resolution portion 1310 and the low image resolution portion 1370 both have a uniform image resolution, the image resolution of the intermediate image resolution portion 1360 can vary depending on a magnitude of a distance from a position in the intermediate image resolution portion 1360 to the position of the gaze point for the predicted image frame 1350. In other words, a number of pixel values per unit area in the portion 1310 in the predicted image frame 1350 is the same throughout the portion 1310 and a number of pixel values per unit area in the portion 1370 is the same throughout the portion 1370, whereas in the intermediate image resolution portion 1360 the number of third pixel values per unit area varies based on a distance from the gaze point. The pixel value density (number of respective pixel values per unit area) is greatest near the boundary with the portion 1310 and smallest near the boundary with the portion 1370, and when traversing the intermediate image resolution portion 1360 in a direction away from the gaze point the pixel density transitions from higher pixel value density to lower pixel value density.

In embodiments of the disclosure, the image resolution of the intermediate image resolution portion 1360 decreases with increasing distance from the gaze point. The intermediate image resolution portion 1360 in the predicted image frame 1350 has a first image resolution at a first position located a first distance from the gaze point in the predicted image frame 1350 (the gaze point corresponds to the centre of the predicted image frame 1350 in Figure 13b) and also has a second image resolution at a second position located a second distance from the gaze point in the predicted image frame 1350, where the first position is located closer to the gaze point than the second position and the first image resolution is higher than the second image resolution. For example, the first position may be located at the boundary between the high image resolution portion 1310 and the intermediate image resolution portion 1360, and the first image resolution may be less than or equal to the image resolution of the high image resolution portion 1310. The second position may be located at the

boundary between the low image resolution portion 1370 and the intermediate image resolution portion 1360, and the second image resolution may be greater than or equal to the image resolution of the low image resolution portion 1370. In some cases, the image resolution of the intermediate image resolution portion 1360 varies linearly with respect to distance from the gaze point such that the mage resolution of the intermediate image resolution portion 1360 decreases linearly with increasing distance from the gaze point.

Figure 14a schematically illustrates a graph showing the change in image resolution for the predicted image frame 1350. In the example shown in Figure 14a, the image resolution of the intermediate image resolution portion 1360 varies linearly with respect to distance from the gaze point. The line A indicates the boundary between the high image resolution portion 1310 and the intermediate image resolution portion 1360, and the line B indicates the boundary between the low image resolution portion 1370 and the intermediate image resolution portion 1360. The high image resolution portion 1310 has a first image resolution R1 and the low image resolution region 1370 has a second image resolution R2, the first image resolution R1 being higher than the second image resolution R2. In the example shown in Figure 14a, the intermediate image resolution portion 1360 has an image resolution that is substantially the same as the first image resolution R1 at (proximate to) the boundary A and has an image resolution that is substantially the same as the second image resolution R2 and the image resolution of the intermediate image resolution portion 1360 varies linearly as a function of distance from the gaze point. However, in some cases intermediate image resolution portion 1360 may have an image resolution that is less than the first image resolution R1 at (proximate to) the boundary A and may have an image resolution that is greater than the second image resolution R2 at (proximate to) the boundary B.

In embodiments of the disclosure, the image resolution of the intermediate image resolution portion 1360 varies non-linearly in dependence upon the distance from the gaze point. Figure 14b schematically illustrates another graph showing the change in image resolution for the predicted image frame 1350, in which the image resolution of the intermediate image resolution portion 1360 varies non-linearly with respect to distance from the gaze point. It will be appreciated that the image resolution of the portion 1360 may vary non-linearly in a manner different to that shown in the example in Figure 14b. For example, whilst Figure 14b shows the image resolution for the portion 360 decreasing most sharply near the boundary A, in some cases the image resolution for the portion 360 may steadily decrease near the boundary A and then decrease most sharply near the boundary B. However, the image resolution for the portion 1360 preferably varies non-linearly in the manner shown in the schematic in Figure 14b so that the image resolution decreases most sharply near the boundary A and then has a more steady decrease near the boundary B, because in this case the drop in image resolution of the intermediate image resolution portion 1360 varies with respect to distance from the gaze point

in a manner that approximates the change in visual resolution acuity as shown in Figure 11. Note that the shape of the curve shown in Figure 11 approximately matches the shape of the curve shown between the boundaries A and B in Figure 14b while allowing a smooth transition of the image resolution at the boundaries.

In embodiments of the disclosure, the intermediate image resolution portion has an image resolution substantially the same as the high image resolution portion at the boundary with the high resolution portion and has an image resolution substantially the same as the low image resolution portion at the boundary with the low resolution portion. As shown in Figure 14b, at (or proximate to) the boundary A the intermediate image resolution portion 1360 has an image resolution that is substantially the same as the image resolution R1 and at (or proximate to) the boundary B the intermediate image resolution portion 1360 has an image resolution that is substantially the same as the image resolution R2.

In embodiments of the disclosure, a total number of respective pixel values in the predicted image frame is greater than a total number of the respective first pixel values in the input image frame. In the case where the predicted image frame comprises the high image resolution portion 1310 and the low image resolution portion 1320, as shown in Figure 13a, the predicted image frame 1300 comprises a first number of respective pixel values including: the second pixel values in the portion 1310; and the first pixel values in the portion 1320, in which only some of the first pixel values obtained by the processing circuitry 1210 for the input image frame are included in the predicted image frame 1300. The pixel value density (number of respective pixel values per unit area) in the portion 1320 is substantially the same as the first pixel value density in the input image frame, whereas the pixel value density in the portion 1310 is greater than the pixel value density in the input image frame. Consequently, the total number of respective pixel values is greater for the predicted image frame 1300. In some examples, the predicted image frame 1300 may be processed so as to remove one or more pixel values from the low image resolution portion 1320. For example, a vignette effect may be applied at the periphery of the predicted image frame 1300 or a portion of the predicted image frame 1300 may be cropped thereby resulting in a reduction in the total number of respective pixel values in the predicted image frame 1300.

Similarly, in the case where the predicted image frame comprises the high image resolution portion 1310, the intermediate image resolution portion 1360 and the low image resolution portion 1370, as shown in Figure 13b, the predicted image frame 1350 comprises a first number of respective pixel values including: the second pixel values in the portion 1310; the third pixel values in the portion 1360; and the first pixel values in the portion 1370, in which only some of the first pixel values obtained by the processing circuitry 1210 for the input image frame are included in the predicted image frame 1350. The number of respective pixel values in the portion 1310 is greater than a number of respective pixel values in a same sized portion in the

input image frame (in other words a pixel values density in the portion 1310 is greater than a pixel value density in the corresponding portion in the input image frame), the number of respective pixel values in the portion 1360 is greater than a number of respective pixel values in a same sized portion in the input image frame, and the number of respective pixel values in the portion 1370 is substantially the same as a number of respective pixel values in a same sized portion in the input image frame. Therefore, the total number of respective pixel values is greater for the predicted image frame 1350 than for the input image frame.

Training Data

The machine learning model 1230 is trained with training image frames each comprising at least a first portion having a first image resolution and a second portion having a second image resolution, the first image resolution being higher than the second image resolution. The training image frames can be generated as part of an execution of a video application that generates images according to a user's gaze point (e.g. a video game and/or a film and/or a recorded television programme) such that each training image frame includes at least a high resolution portion centred upon the gaze point and a low resolution portion corresponding to a peripheral region. For example, as part of an execution of an application such as a computer game, image frames can be generated and at least some of the image frames can be acquired and used as training image frames for use in training the machine learning model 1230. In this case, it may be possible to directly acquire training image frames having a high image resolution portion and a low image resolution portion from an executing application. Optionally, it may be possible to acquire the gaze data corresponding to the image frames generated by executing the application. Alternatively, an estimate of the gaze point may be made for each image frame based on an assumption that the gaze point corresponds to a centre of the high resolution portion in the image frame. In some cases, the training image frames can be generated for a given video content item (e.g. a given video game or a given film) during a development phase or quality assurance testing phase of the video content item and used to train the machine learning model 1230 for the given video content item.

Alternatively, an image frame previously generated for a given video content can be acquired and in the case where the image frame has a single image resolution (a high resolution image frame such as a 4k image frame, for example), the second portion of the image frame can be processed so as to cull pixel values to decrease an image resolution in the second portion, thereby achieving a training image frame having a high image resolution portion and a low image resolution portion. In this case, an estimated gaze point may be obtained for each image frame by calculating a point corresponding to the centre of the high resolution portion of the image frame.

A low resolution image frame can be generated for each training image frame by downsampling a training image frame so as to obtain a low resolution image frame for the

training image frame. Therefore, a training image frame can be processed so as to reduce the image resolution and obtain a low resolution image frame having a smaller number of respective pixel values and a given image resolution. The training image frames can each be processed so as to obtain a corresponding lower resolution image frame having any suitable image resolution. Preferably, the lower resolution image frames will have an image resolution that is either the same as (or similar to) an image resolution of the image frames for which the first pixel values are obtained by the processing circuitry 1210.

The machine learning model 1230 can thus be trained to learn a function for outputting the training image frame from an input including the low resolution image frame and a gaze point, where the input gaze point corresponds to the centre of the high resolution portion of the training image frame. Therefore, the machine learning model 1230 can be trained to output the training image frame having at least the high resolution portion and the low resolution portion for an input low resolution image frame, where the high resolution portion in the training image frame has a number of respective pixel values that is greater than a number of respective pixels in the same portion in the low resolution image frame. The machine learning model 1230 can be trained with a plurality of such training image frames, such that when provided with a gaze point and an input image frame comprising first pixel values, the machine learning model 1230 is trained to predict second pixel values for at least a portion of the input image frame so as to increase the image resolution for the portion of the input image frame and improve the image resolution for the portion of the input image frame. The training image frames are a ground truth for the machine learning model 1230, in that the training image frames represent examples of ideal predicted image frames that are to be output for an input image frame.

As discussed previously, in some embodiments of the disclosure the training image frames each comprises a third portion between the first portion and the second portion, the third portion corresponding to a transition region and having an image resolution higher than the second image resolution and lower than the first image resolution. In this case, the training image frames can be generated as part of an execution of a video application that generates images according to a user's gaze point (e.g. a video game and/or a film and/or a recorded television programme) such that each training image frame includes the first portion (high resolution), the second portion (low resolution) and the third portion (intermediate resolution) which each have a uniform image resolution. In some cases the training image frames having the first, second and third portions may be generated by a system performing multiple render passes to allow the image frame to be rendered multiple times at different image resolutions and the resulting renders are then composited together to achieve respective portions of different image resolution. Alternatively, in some cases certain hardware can be used that allows rendering at different resolutions in different parts of an image frame without needing additional render passes to achieve respective portions of different image resolution. Therefore,

in some examples, the training image frame may comprise a third region having a uniform image resolution.

However, as explained previously the use of such hardware can result in limitations as to the smoothness of the transition between regions of different image resolution and may only be suitable for obtaining a small number of regions each having a uniform image resolution. Training image frames obtained in this way can be post-processed so that the third portion in the training image frame is processed in accordance with a smoothing function so that the image resolution varies with respect to distance from the centre of the first portion. For example, post-processing may be performed for a training image frame generated in this way to selectively cull pixel values according to distance from the centre of the first portion. In this way, training image frames can be obtained having the first, second and third portions, in which the image resolution of the third portion varies in dependence upon a distance from the centre point of the first portion either linearly or non-linearly. Therefore, training image frames having an image resolution that varies with respect to distance from the centre point of the first portion (high resolution portion) in the training image frame in a manner similar to that shown in the schematic in Figure 14a and Figure 14b can be obtained.

Alternatively, an image frame previously generated for a given video content can be acquired and in the case where the image frame has a single image resolution (a high resolution image frame such as a 4k image frame, for example), the third portion and the second portion of the image frame can each be processed so as to cull pixel values to decrease an image resolution in the third portion and the second portion, thereby achieving a training image frame having the first, second and third portions, in which the image resolution of the third portion varies in dependence upon a distance from the centre point of the first portion either linearly or non-linearly.

In embodiments of the disclosure, the third portion in each training image frame is generated by applying a post-processing smoothing function to the training image frame. The training image frames can be generated by post-processing image frames generated for a given video content item. A training image frame having a single image resolution (e.g. a 4k image frame) can be post-processed by applying a smoothing function as discussed above so as to obtain the first, second and third portions. Alternatively, an image frame having the first, second and third portions may be generated as part of an execution of an application such that the third portion has a uniform image resolution. The third portion can then be processed in accordance with the smoothing function so that the image resolution for the third portion varies with respect to distance from the centre of the first portion by selectively culling pixel values according to distance from the centre of the first portion so that a pixel value density for the third portion is greater nearer the boundary with the first portion and smaller nearer the boundary with the second portion.

In the case where the image resolution of the third portion of the training image frame is to vary linearly, then the third portion in each training image frame can be generated by applying a post-processing smoothing function that varies linearly with respect to distance. Similarly, in the case where the image resolution of the third portion of the training image frame is to vary non-linearly, then the third portion in each training image frame can be generated by applying a post-processing smoothing function comprising human vision resolution acuity information for the human eye so that the third portion has an image resolution that varies in dependence upon a distance from the centre point of the first portion so as to approximate the change in human visual acuity as shown in Figure 11. In this way, training image frames can be obtained which include the first portion (high resolution), second portion (low resolution) and the third portion (intermediate resolution) having an image resolution that varies either linearly or non-linearly. At the boundary where the third portion and first portion terminate, the part of the third portion adjacent to the boundary can be post-processed using the smoothing function so that there is a gradual change in image resolution. Similarly, at the boundary where the third portion and second portion terminate, the part of the third portion adjacent to the boundary can be post-processed using the smoothing function so that there is a gradual change in image resolution. As such, by post-processing the training image frames using a smoothing function for at least the third portion (optionally the smoothing function may be applied to both the first portion and the third portion to smooth the boundary and may also be applied to both the third portion and the second portion to smooth the boundary), a training image frame can be obtained having an image resolution that varies according to distance from the centre of the first portion in the manner as shown previously in Figure 14b in relation to the predicted image frame 1350. Therefore, training image frames having the varying image resolution as shown in Figure 14b can be obtained for training the machine learning model 1230. The training image frames obtained in this way are a ground truth for the machine learning model 1230 in that the training image frames represent examples of predicted image frames that are to be output for an input image frame.

In embodiments of the disclosure, each training image frame comprises one or more post-processing effects based upon image filtering or image processing or the like, with a non-exhaustive list of examples comprising blurring effects; anti-aliasing; lighting effects; and colour contrast enhancement. The ability of the human eye to discern detail is reduced for peripheral portions of the user's vision, as shown in Figure 11. However, whilst peripheral vision quality is significantly reduced compared to foveal vision quality, certain stimuli in the peripheral region are more noticeable than others, such as motion. Colour perception also degrades with increasing distance from the gaze point, but even for peripheral portions of the user's vision colour differentiation is still possible. Post-processing rendering effects can typically be applied to image frames to adjust an appearance of the image frame to improve a sense of realism.

Therefore, when implementing foveated rendering conventional image rendering systems may typically apply a post-processing blurring effect (e.g. Gaussian blurring function), anti-aliasing and/or colour contrast enhancement for a peripheral portion of an image frame. Image frames including at least the first and second portion (and optionally the third portion) and one or more such post-processing effects can be used as training image frames such that the machine learning model 1230 can be trained using such training image frames as a ground truth. Therefore, for a predicted image frame, such as the predicted image frame 1350 in Figure 13b, fourth pixel values can be predicted for some of the portion 1370 so that the portion 1370 includes one or more such post-processing effects.

In embodiments of the disclosure, the machine learning model 1230 is trained to predict fourth pixel values for a peripheral portion of the image frame not within a threshold distance of the gaze point in dependence upon the first pixel values and the gaze point, the fourth pixel values corresponding to the low image resolution portion 1370 of the predicted image frame 1350 such that the low image resolution portion 1370 of the predicted image frame 1350 includes one or more of the post-processing effects. In addition to predicting the second pixel values for the high image resolution portion 1310 (and optionally the third pixel values for the intermediate image resolution portion 1360), the machine learning model 1230 can optionally be trained to predict the fourth pixel values for the low image resolution portion 1370 of the predicted image frame 1350. Therefore, in some embodiments the low image resolution portion 1370 of the predicted image frame 1350 comprises some of the first pixel values obtained by the processing circuitry 1210 and may also comprise fourth pixel values predicted by the machine learning model 1230.

In embodiments of the disclosure, the image frame for which the first pixel values are input to the machine learning model 1230 and the training image frames used for training the machine learning model 1230 correspond to a same type of video content. For example, the type of video content may be a type of sporting event such as football matches or basketball matches. Alternatively, the type of video content may be a type of movie or may be a type of video game, such as a genre of movie or video game.

For example, the machine learning model 1230 can be trained using a plurality of training image frames of football matches. The data processing system 1200 may be provided as part of a user's personal computer or game console and configured to receive an encoded stream of image data for a football match. The processing circuitry 1200 can be configured to decode the encoded stream and obtain first pixel values for each image frame in the stream. The gaze point of the user viewing one or more previous image frames output for display to the user by the system 1200 can be detected and an input comprising information indicative of the gaze point can be received by the input circuitry 1220. The first pixel values for an image frame and the gaze point can therefore be input to the machine learning model 1230, and second pixel

values (and optionally third pixel values, and optionally fourth pixel values) can be predicted as described previously so as to obtain a predicted image frame for the input image frame such that the predicted image frame includes at least a portion for which the image resolution is enhanced. Therefore, for cases in which network conditions result in image frames with a reduced image resolution being received by a user's device, the data processing system 1200 can output predicted image frames for the sequence of low image resolution input frames, in which the image resolution in a portion of each predicted image frame corresponding to the gaze direction is enhanced.

In some cases, the image frame for which the first pixel values are input to the machine learning model 1230 and the training image frames used for training the machine learning model 1230 correspond to same video content item (e.g. same video game). For example, a plurality of training image frames can be acquired for a given video content item, such as given video game (e.g. Ratchet and Clank) so as to train the machine learning model 1230 for predicting second pixel values (and optionally third pixel values and optionally fourth pixel values) for obtaining predicted image frames for the given video content item when low resolution image frames are received by the data processing system 1200 for the video content item. In this way, the machine learning model 1230 can be trained for a specific video content item.

In embodiments of the disclosure, the image frame for which the first pixel values are input to the machine learning model 1230 and the training image frames used for training the machine learning model 1230 correspond to a same type of scene. The machine learning model 1230 may be trained using training image frames for a given type of scene (e.g. urban landscape), in which the training image frames may correspond to one or more different video content items (e.g. one or more different video games) or may correspond to a same video content item (e.g. a same video game). In some cases, the machine learning model 1230 may be trained using training images for a given type of scene in a given video content item. For example, the machine learning model 1230 may be trained for urban landscape scenes in a given video game title such as the video game Call of Duty: Modern Warfare.

A first machine learning model 1230 may be trained for a first type of scene in general and a second machine learning model 1230 may be trained for a second type of scene in general. Alternatively, a first machine learning model 1230 may be trained for a first type of scene in a given video content item and a second machine learning model 1230 may be trained for a second type of scene in the same video content item. In some cases, the machine learning model 1230 can be trained for a specific level in a video game. For instance, a first machine learning model 1230 may be trained for one level in a video game and a second machine learning model 1230 may be trained for another level in the video game, in which a type of scene differs for the two different levels.

The data processing system 1200 may comprise more than one machine learning model 1230. In some cases, the data processing system 1200 comprises a plurality of machine learning models 1230a, 1230b ... 1230n, in which each machine learning model is trained for a type of scene. For example, a first machine learning model 1230a can be trained with training image frames for a first type of scene (e.g. countryside scene) and a second machine learning model 1230b can be trained with training image frames for a second type of scene (e.g. urban landscape scene). It will be appreciated that training image frames can be acquired for various types of scenes. In this way, the processing circuitry 1210 can obtain the first pixel values for an image frame corresponding to a particular type of scene (e.g. ocean scene) and the first pixel values and the gaze point can be provided as an input to a machine learning model 1230 trained using training images for that type of scene. Therefore, in some embodiments the data processing system 1200 comprises two or more machine learning models, in which one machine learning model is trained to predict the second pixel values for image frames having a first type of scene and the other machine learning model is trained to predict the second pixel values for image frames having a second type of scene.

In some examples, the data processing system 1200 is configured to receive an encoded stream of image data in which each image frame has associated information indicating a scene type for the image frame. For example, a content server may associate such information with an image frame. Alternatively, in some examples the data processing system 1200 may comprise an image classification machine learning model (not shown in Figure 12) trained to assign a classification to an image frame. The image classification machine learning model may be trained using known techniques for performing machine learning based image classification. In response to the assigned classification, the first pixel values for each image frame can be provided as an input to one of the machine learning models from the plurality of machine learning models 1230a, 1230b ... 1230n so that an image frame having a given scene type is input to a machine learning model trained for that given scene type.

In embodiments of the disclosure, the image frame for which the first pixel values are input to the machine learning model 1230 and the training image frames used for training the machine learning model 1230 correspond to a same video content item, in which the training image frames are selected from candidate training image frames for the video content item in dependence upon gaze information recorded for a plurality of users for the video content item. A plurality of training image frames can be acquired for a given video content item, such as given video game, so as to train the machine learning model 1230 for predicting second pixel values (and optionally third pixel values and optionally fourth pixel values) to obtain predicted image frames for the given video content item (or for a given scene within the given video content item). The plurality of training image frames used for training the machine learning model 1230 can be selected from a plurality of candidate training image frames for the video content item,

such that the training image frames selected for use in training the model 1230 correspond to a subset of the candidate training image frames.

A plurality of candidate training image frames can be recorded for at least a first user's interaction with the video content item and corresponding gaze information for the first user can also be recorded. For each of a plurality of users for the video content item, both the image frames generated for display to each user and the gaze points detected for each user can be recorded. As such, a candidate training image frame database can be created for the video content item comprising the candidate training image frames generated for display to each of the users.

The gaze information for each user may comprise a plurality of gaze points detected for the user while viewing the image frames, each gaze point having an associated timestamp. The gaze information recorded for the plurality of users of the video content item can be analysed to generate gaze heatmap information for the video content item indicative of where the plurality of users look at most within the different parts of the video content item. For example, for a video game, gaze heatmap information for the environment within the video game may be generated for each user and by combining the gaze heatmap information for the plurality of users a heatmap can be obtained indicating parts of the environment looked at most by the plurality of users. Using the gaze heatmap information, a subset of the candidate training image frames recorded in the database for the plurality of users can be selected so that candidate training image frames corresponding to parts of the video content item looked at most by the users can be selected and the selected candidate training image frames can be used for training the machine learning model 1230. In this way, the training image frames can be biased so as to focus most on where users look at the most within the different parts of the game, thereby allowing the machine learning model 1230 to be trained in a more focussed way for a given video content item, or even a given level or scene within a video content item.

Referring now to Figure 15, in embodiments of the disclosure a data processing method comprises:

obtaining (at a step 1310) first pixel values for an image frame;

receiving (at a step 1320) an input indicative of a gaze point of an eye of a user for the image frame;

inputting (at a step 1330) the first pixel values and the gaze point to a machine learning model trained to predict second pixel values for a portion of the image frame in dependence upon the first pixel values and the gaze point, the machine learning model trained with training image frames each comprising a first portion having a first image resolution and a second portion having a second image resolution, the first image resolution being higher than the second image resolution;

predicting (at a step 1340), by the trained machine learning model, the second pixel values for the portion of the image frame, a number of the respective second pixel values predicted for the portion of the image frame being greater than a number of the respective first pixel values obtained for the portion of the image frame;

storing (at a step 1350) a predicted image frame comprising the second pixel values and some of the first pixel values, the second pixel values corresponding to a high image resolution portion of the predicted image frame; and

outputting (at a step 1360) the predicted image frame for display to the user.

It will be appreciated that example embodiments can be implemented by computer software operating on a general purpose computing system such as a games machine. In these examples, computer software, which when executed by a computer, causes the computer to carry out any of the methods discussed above is considered as an embodiment of the present disclosure. Similarly, embodiments of the disclosure are provided by a non-transitory, machine-readable storage medium which stores such computer software.

Thus any required adaptation to existing parts of a conventional equivalent device may be implemented in the form of a computer program product comprising processor implementable instructions stored on a non-transitory machine-readable medium such as a floppy disk, optical disk, hard disk, solid state disk, PROM, RAM, flash memory or any combination of these or other storage media, or realised in hardware as an ASIC (application specific integrated circuit) or an FPGA (field programmable gate array) or other configurable circuit suitable to use in adapting the conventional equivalent device. Separately, such a computer program may be transmitted via data signals on a network such as an Ethernet, a wireless network, the Internet, or any combination of these or other networks.

It will also be apparent that numerous modifications and variations of the present disclosure are possible in light of the above teachings. It is therefore to be understood that within the scope of the appended claims, the disclosure may be practised otherwise than as specifically described herein.

CLAIMS

1. A data processing system comprising:

processing circuitry to obtain first pixel values for an image frame;

input circuitry to receive an input indicative of a gaze point of an eye of a user for the image frame;

a machine learning model trained to predict second pixel values for a portion of the image frame in dependence upon the first pixel values and the gaze point, the machine learning model trained with training image frames each comprising a first portion having a first image resolution and a second portion having a second image resolution, the first image resolution being higher than the second image resolution, in which a number of the respective second pixel values predicted for the portion of the image frame is greater than a number of the respective first pixel values obtained by the processing circuitry for the portion of the image frame;

storage circuitry to store a predicted image frame comprising the second pixel values and some of the first pixel values obtained by the processing circuitry for the image frame, in which the second pixel values correspond to a high image resolution portion of the predicted image frame; and

output circuitry to output the predicted image frame for display to the user;

in which the image frame and the training image frames correspond to a same video content item, and in which the training image frames are selected from candidate training image frames for the video content item in dependence upon a heatmap generated from gaze information recorded for a plurality of users for the video content item, so that candidate training image frames corresponding to parts of the video content item looked at most by the plurality of users can be selected.

2. The data processing apparatus according to claim 1, in which the machine learning model is trained to predict third pixel values for another portion of the image frame in dependence upon the first pixel values and the gaze point, in which the another portion of the image frame at least partially surrounds the portion of the image frame, the training image frames each comprising a third portion between the first portion and the second portion, the third portion corresponding to a transition region and having an image resolution higher than the second image resolution and lower than the first image resolution.

3. The data processing apparatus according to claim 2, in which a number of the respective third pixel values predicted for the another portion of the image frame is greater than a number of the respective first pixel values obtained by the processing circuitry for the another portion of the image frame.

4. The data processing apparatus according to claim 2 or claim 3, in which the third pixel values in the predicted image frame correspond to an intermediate image resolution portion of the predicted image frame, and the first pixel values in the predicted image frame correspond to a low image resolution portion of the predicted image frame.

5. The data processing apparatus according to claim 4, in which the image resolution of the intermediate image resolution portion varies in dependence upon a distance from the gaze point, and the image resolution of the third portion of each training image frame varies in dependence upon a distance from a centre point of the first portion.

6. The data processing apparatus according to claim 5, in which the image resolution of the intermediate image resolution portion varies non-linearly in dependence upon the distance from the gaze point.

7. The data processing apparatus according to claim 5 or claim 6, in which the image resolution of the intermediate image resolution portion decreases with increasing distance from the gaze point.

8. The data processing apparatus according to any one of claims 5 to 7, in which the intermediate image resolution portion has an image resolution substantially the same as the high image resolution portion at the boundary with the high resolution portion and has an image resolution substantially the same as the low image resolution portion at the boundary with the low resolution portion.

9. The data processing apparatus according to any one of claims 2 to 8, in which the third portion in each training image frame is generated by applying a post-processing smoothing function to the training image frame.

10. The data processing apparatus according to any preceding claim, in which each training image frame comprises one or more post-processing effects based upon image filtering or processing.

11. The data processing apparatus according to claim 10, in which the machine learning model is trained to predict fourth pixel values for a peripheral portion of the image frame not within a threshold distance of the gaze point in dependence upon the first pixel values and the gaze point, the fourth pixel values corresponding to a low image resolution portion of the

predicted image frame such that the low image resolution portion of the predicted image frame includes one or more of the post-processing effects.

12. The data processing apparatus according to any preceding claim, in which the training image frames are a ground truth for the machine learning model.

13. The data processing apparatus according to any preceding claim, in which a total number of respective pixel values in the predicted image frame is greater than a total number of the respective first pixel values in the image frame.

14. The data processing apparatus according to any preceding claim, in which processing circuitry is configured to obtain the first pixel values for each of a plurality of image frames and the output circuitry is configured to output a predicted image frame for each of the image frames.

15. The data processing apparatus according to any preceding claim, in which the processing circuitry comprises either rendering circuitry to perform rendering operations to obtain the first pixel values or decoding circuitry to perform decoding operations to obtain the first pixel values.

16. The data processing apparatus according to any preceding claim, in which the image frame and the training image frames correspond to a same type of video content.

17. The data processing apparatus according to any preceding claim, in which the image frame and the training image frames correspond to a same type of scene.

18. The data processing apparatus according to any preceding claim, comprising another machine learning model, in which the machine learning model is trained to predict the second pixel values for image frames having a first type of scene and the another machine learning model is trained to predict the second pixel values for image frames having a second type of scene.

19. A data processing method, comprising:

obtaining first pixel values for an image frame;

receiving an input indicative of a gaze point of an eye of a user for the image frame;

inputting the first pixel values and the gaze point to a machine learning model trained to predict second pixel values for a portion of the image frame in dependence upon the first pixel values and the gaze point, the machine learning model trained with training image frames each

comprising a first portion having a first image resolution and a second portion having a second image resolution, the first image resolution being higher than the second image resolution;

predicting, by the trained machine learning model, the second pixel values for the portion of the image frame, a number of the respective second pixel values predicted for the portion of the image frame being greater than a number of the respective first pixel values obtained for the portion of the image frame;

storing a predicted image frame comprising the second pixel values and some of the first pixel values, the second pixel values corresponding to a high image resolution portion of the predicted image frame; and

outputting the predicted image frame for display to the user;

in which the image frame and the training image frames correspond to a same video content item, and in which the training image frames are selected from candidate training image frames for the video content item in dependence upon a heatmap generated from gaze information recorded for a plurality of users for the video content item, so that candidate training image frames corresponding to parts of the video content item looked at most by the plurality of users can be selected.

20. Computer software which, when executed by a computer, causes the computer to perform the method of claim 19.