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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
26 January 2023 (26.01.2023)

WIPO | PCT

(10) International Publication Number
WO 2023/001816 A1

(51) International Patent Classification:
G06T 19/00 (2011.01)

(21) International Application Number:
PCT/EP2022/070186

(22) International Filing Date:
19 July 2022 (19.07.2022)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/223,182 19 July 2021 (19.07.2021) US

(71) Applicant: **SCANERGY GMBH** [CH/CH]; Hallauerstrasse 30, 8213 Neunkirch (CH).

(72) Inventors: **HÖNGER, Josua**; c/o Scanergy GmbH, Hallauerstrasse 30, 8213 Neunkirch (CH). **ANGELOV, Zoran**; c/o Scanergy GmbH, Hallauerstrasse 30, 8213 Neunkirch (CH). **ROSSI, Markus**; Rütiwiesstrasse 24, 8645 Jona (CH).

(74) Agent: **FREI PATENT ATTORNEYS (ZUSAMMENSCHLUSS 214)**; c/o Frei Patentanwaltsbüro AG, Postfach, 8032 Zurich (CH).

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

(54) Title: AUGMENTED REALITY VIRTUAL CAMERA

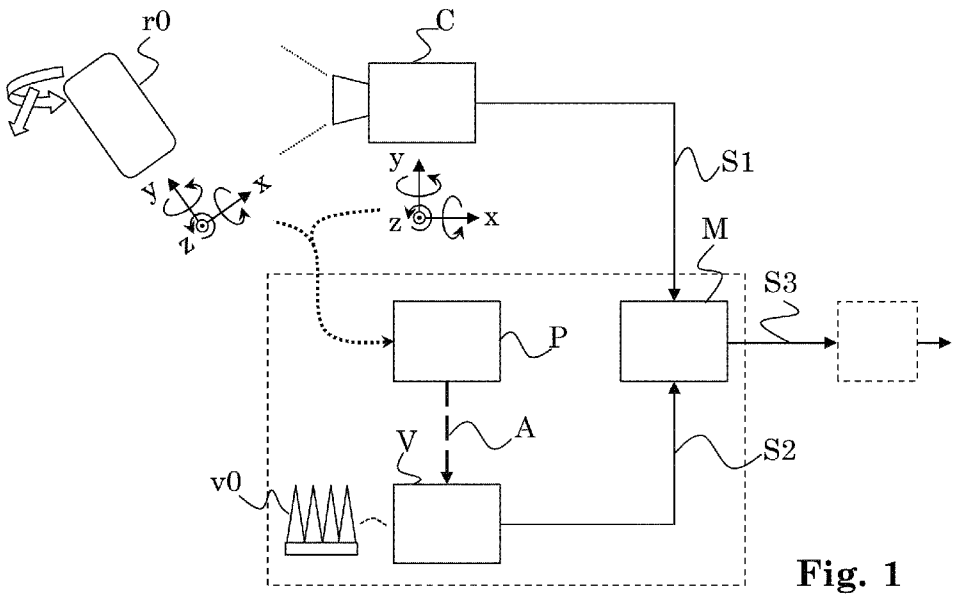

Fig. 1

(57) Abstract: The method for generating an augmented video stream (S3), comprises providing a real video camera (C); generating a first video stream (S1) comprising real-time video stream data from the real video camera (C); providing a real object (rO); determining in real-time relative pose data (A) indicative of a relative position in space and of a relative orientation in space of the real video camera (C) and the real object (rO); generating a second video stream (S2) comprising a representation of a virtual object (vO); modulating the representation of the virtual object (vO) in the second video stream (V2) in real-time and in dependence of the relative pose data (A); and outputting and/or generating the augmented video stream (S3) from the first video stream (SI) and the second video stream (S2) in real-time. This way, a pose of the virtual object (vO) can be controllable by moving the real object (rO).

WO 2023/001816 A1

[Continued on next page]

TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

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

AUGMENTED REALITY VIRTUAL CAMERA

The invention relates to augmented reality technology. It relates to methods and apparatuses according to the opening clauses of the claims.

A wealth of technologies for generating an augmented video stream are known in the art.

The inventors contemplated a new way of generating an augmented video stream which enables simple and intuitive interaction by a user. In particular, they contemplated a way which makes possible a simple and intuitive modulation in real-time of a representation of a virtual object. Moreover, some embodiments of the invention can be

carried out using current standard hardware, such as modern smartphones and laptop computers, plus dedicated software. The invention can furthermore find application in video conferencing, namely the augmented video stream can be used as one participant's video signal in today's video conferencing software and thus be transmitted to further participants.

Further objects and various advantages emerge from the description and embodiments below.

At least one of these objects is at least partially achieved by systems and methods according to the patent claims.

In particular, the method for generating an augmented video stream can comprise

- providing a real video camera;
- generating a first video stream comprising real-time video stream data from the real video camera;
- providing a real object;
- determining in real-time relative pose data indicative of a relative position in space and of a relative orientation in space of the real video camera and the real object;
- generating a second video stream comprising a representation of a virtual object;
- modulating the representation of the virtual object in the second video stream in real-time and in dependence of the relative pose data;
- outputting and/or generating from the first video stream and the second video stream the augmented video stream in real-time.

And, correspondingly, the system (or combination) for generating an augmented video stream can comprise

— a first video unit comprising a real video camera, configured to generate a first video stream comprising real-time video stream data from the real video camera;

— a real object;

— a sensing unit configured to determine in real-time relative pose data indicative of a relative position in space and of a relative orientation in space of the real video camera and the real object;

— a second video unit operationally connected to the sensing unit for receiving the relative pose data from the sensing unit and configured

  – to generate a second video stream comprising a representation of a virtual object; and

  – to modulate the representation of the virtual object in the second video stream in real-time and in dependence of the relative pose data;

— a video processing unit operationally connected to the first video unit and to the second video unit, configured to output and/or generate from the first video stream and the second video stream the augmented video stream in real-time.

The invention can make possible that the representation of the virtual object (more particularly: its representation in the augmented video stream) can be modulated in real-time by moving the real object. And it is possible to accomplish this in a synchronized manner. Accordingly, a user can modify the augmented video stream and more particulary the representation of the virtual object therein in a very simple and intuitive way. Movements of the real object can be translated into (virtual) movements of the virtual object in the augmented video stream. The real object can be used by a user as a controller for controlling (virtual) movements of the virtual object in the augmented video stream. For example, the invention can enable to create the illusion that the virtual object in the augmented video stream moves in real space exactly as a real-world item firmly connected to the real object would do.

The real video camera is a physically existing video camera; in contrast to a merely virtual camera. Thus, it will typically comprise a photo sensor, such as semiconductor RGB imaging chip, and usually also one or more optical elements, such as one or more lenses. In the context of the instant invention, it is, at least in most embodiments, unnecessary to move the real video camera, i.e. it can remain in one and the same position and orientation in space, i.e. remain in one and the same pose (the combination of position and orientation of an object is referred to as the "pose" of an object).

Typically, the real video camera is configured to image a (real; physically existing) scene within its field of view, typically in proximity to the real video camera.

In some embodiments, the real video camera is a 2D video camera, such as a camera generating a sequence of color images.

In some embodiments, the real video camera is a 2D-plus-depth video camera, such as a camera generating a sequence of color images (in one layer) containing (in another layer) depth information, i.e. at least a portion of the pixels bear information regarding a distance between the camera and an object imaged at the respective pixel. Such cameras are known in the art. They can be based, e.g., on image processing and machine learning, such that the real video camera may comprise a processing unit for these purposes; and/or they can be based on stereo imaging, such that the real video camera may comprise two cameras and a processing unit for this purpose; and/or they can be based on other techniques, such as time-of-flight sensing, structured light imaging, LiDAR (light detection and ranging) in which cases the real video camera can comprise a light emitter, such as an infrared light source for such purposes, e.g., emission of structured light.

In some embodiments, the real video camera is a 3D video camera (volumetric camera), such as a camera generating a sequence of volumetric data. Such a video camera can comprise for this purpose, e.g., a processing unit and a plurality of subordinate video cameras viewing a scene to be imaged from different positions and/or angles.

In some embodiments, the real video camera is a peripheral camera of a computing device, such as a camera with wireless, e.g., "Bluetooth"-based or wirebound, e.g., USB-based interconnectivity.

In some embodiments, the real video camera is a built-in camera of a computing device, such as of smartphone, a laptop computer, a tablet computer, a desktop computer – more precisely of a monitor of the desktop computer.

A video stream, and in particular the first video stream and/or the second video stream and/or the augmented video stream, is a sequence (more particularly: sequence in time) of frames. Typically, such a frame comprises image data, e.g., it can constitute image data. In many cases, the image data are 2D color data, such as data describing RGB pixels. It is also possible that such a frame comprises volumetric data, such as a stack of image data or differently defined data describing properties at voxels ("3D pixels") such as color information at grid points of a 3D grid in a volume.

Furthermore, such a frame can, in instances, comprise more than one data layer. More particularly, a frame can comprise, in addition to said image data (constituting a data layer) or to said volumetric data (constituting a data layer), one or more additional data layers. Such additional data layers can comprise (and in particular constitute), e.g., depth image data (indicative of distances along a depth direction), confidence level data (indicative of a reliability or trustability of data of another layer), meta data, such as pose data of an object in the frame.

A frame and, more particularly each layer of a frame, can optionally comprise one or more (fully) transparent pixels (or voxels), i.e. pixels (or voxels) bearing no information (such as no color information, no depth information, no confidence information).

Size and shape of a frame is not particularly limited, e.g., it does not need to be contiguous or does not need to be rectangular, which applies also simple color image frames. For example, a color image frame, e.g., of the second video stream, may show a representation (or view) of a (e.g., small round) virtual object only; and alternatively, e.g., also as a possibility for the second video stream, a frame having this contents can

be contiguous and rectangular, namely by comprising, as mentioned above, transparent pixels: The frame, e.g., showing a representation (or view) of a (e.g., small round) virtual object, while all other pixels are (fully) transparent.

In some embodiments, the video streams (first video stream, second video stream, augmented video stream) are in an uncompressed data format. This can improve performance and thus facilitate the real-time processing. However, the video streams could also be in a compressed data format.

In some embodiments, the first video stream consists of real-time video stream data from the real video camera. E.g., the first video stream is simply the (unaltered) output of the real video camera.

The first video unit can be identical to the real video camera, e.g., in this case.

However, in some embodiments, the first video stream is obtained by altering, e.g., processing the real-time video stream data from the real video camera. E.g., the first video stream can comprise merely a portion of the real-time video stream data; or the first video stream can comprise further video information, such as virtual contents, e.g., by replacing a portion of the real-time video stream data by virtual contents.

The first video unit can comprise further real and/or virtual video cameras and/or a processing unit, e.g., in this case.

The real object is a physically existing object – in contrast to a virtual object. In particular, it can be a movable object, particularly in the sense that its size and weight are such that it can be readily moved by an average human being, e.g., using a hand only.

In some embodiments, the real object is a part of a human body, in particular a hand or a part of a hand. This enables simple and intuitive operation.

In some embodiments, the real object is a hand-held device.

In some embodiments, the real object is an office supplies item such as a writing utensil.

In some embodiments, the real object is a handheld computing device, such as a smartphone or a tablet computer.

In some embodiments, the real object is an add-on-device for such a handheld computing device, in particular attached to the handheld computing device.

In some embodiments, the real object is a device comprising one or more components of the sensing unit, in particular one or more sensors of the sensing unit, e.g., the device can comprise one or more sensors for sensing its position and/or its orientation in space. Said device can be, e.g., a handheld computing device as mentioned above, said sensor being, e.g., a built-in sensor of the handheld computing device. In another example, said device can be, e.g., an add-on device as mentioned above, said one or more sensors being built-in sensors of the add-on device.

The relative pose data could also be referred to as or considered "arrangement data", as they describe the relative arrangement (in space) of the real object and the real video camera.

In particular, the term "relative" pose/position/orientation of (or between) the real object and the real video camera does not specify whether it is, e.g., a pose/position/orientation of the real object with respect to the real video camera, or a pose/position/orientation of the real video camera with respect to the real object. It can can be, e.g., any of these.

There are various ways to determine the relative pose data (a "pose" describing both, position and orientation).

In a first example, for both, the real video camera and the real object, the respective pose is determined, relative to one and the same coordinate system, and from this, the relative pose is determined.

In a second, but similar example, the pose of the real video camera is determined in one coordinate system, and the pose of the real object is determined in another coordinate system different from the first coordinate system. Then, the two coordinate systems are

interrelated, e.g., by a calibration procedure, and from this, finally the relative pose is determined.

For determination of a pose (or at least of a position in space or an orientation in space) of an object, such as of the real object or the real video camera, several techniques are known and available, such as

- accelerometric techniques;
- gyroscopic techniques;
- gravity-based techniques;
- techniques based on determination of the environment of the object, which can comprise, e.g., techniques based on
    - algorithmic evaluation of video data;
    - image analysis with object recognition;
    - machine-learning supported evaluation of video data,
    - deep learning supported evaluation of video data,
    - artificial intelligence-based evaluation of video data;
    - depth-sensing techniques (e.g., based on time-of-flight sensing, based structured light imaging, based on stereo imaging, LiDAR);
- combinations of two or more of these.

Commercially available sensor combinations which could be used, comprise, e.g., sensor combination associated with augmented reality development toolkits such as "ARKit" (by Apple), "ARCore" (by Google), "Vuforia" (by PTC), or sensors of "Kinect" (by Microsoft).

Accordingly, the sensing unit can comprise one or more sensors, such as sensors or sensor combinations as mentioned above and, usually also a processing unit for processing, e.g., interrelating data and/or evaluating data, such as converting sensor raw data into calibrated data.

In some embodiments, the receiving by the second video unit of the relative pose data from the sensing unit is accomplished in a wireless fashion. Correspondingly, the

relative pose data can be transmitted from the sensing unit to the second video unit in a wireless fashion.

In some embodiments, the sensing unit is distributed over two or more devices. Communication between the devices for exchange of data related to sensing results obtained by at least one sensor of the sensing unit can be accomplished in a wireless fashion, in particular if a portion of the sensing unit, e.g., a sensor, is comprised in the real object. Accordingly, the sensing unit can comprise a wireless communication capability, e.g., embodied as communication units in the respective devices.

For example, a portion of the sensing unit can be comprised in real object, and another portion in another device such as in a computing device, e.g., in a computing device comprising the real video camera.

The second video unit can be embodied in form of software implmented in a computing device, in particular in a graphics processing unit (GPU) of the computing device. Plenty suitable computer applications are known and available enabling generation of a (second) video stream comprising a representation of a virtual object and enabling to modulate the representation of a virtual object in the (second) video stream in real-time and in dependence of pose data.

The video processing unit can be embodied in form of software running on a computing device, in particular in a graphics processing unit (GPU) of the computing device. Plenty suitable computer applications are known and available enabling a real-time generation of a video stream (the augmented video stream) from two video streams, such as from the first video stream and the second video stream.

The video processing unit can be, e.g., a video mixer.

The augmented video stream typically comprises data derived from the first video stream and data derived from the second video stream. E.g., the augmented video stream (in a simple example) can comprise at least a portion of the first video stream and at least a portion of the second video stream.

Generating the augmented video stream can comprise merging the representation of the virtual object into the first video stream.

The augmented video stream can be generated, e.g., in a frame-wise manner.

For example, generating the augmented video stream can comprise repeatedly (frame-by-frame) grabbing a frame of the first video stream (first frame) and a (simultaneous) frame of the second video stream (second frame) and creating a new frame from those two frames, which then constitutes a frame of the augmented video stream (augmented frame). For creating the augmented frame, the two frames can be merged, overlayed or be otherwise combined. For example, assuming that the second video stream (and second frame) shows a representation (or view) of the virtual object, while all other pixels (or voxels) are (fully) transparent, the augmented frame can be created by replacing in the first frame those pixels which in the second frame are not fully transparent. (In case of presence of merely partially transparent pixels in the second frame, a respective pixel of the augmented frame could be obtained using both, data from the respective pixel of the second frame and data from the respective pixel of the first frame.)

For example, generating the augmented video stream can be accomplished without separately storing (e.g., in computer memory) the second video stream, not even for a single frame. Namely by generating, e.g., frame-by-frame, the data representing the representation of the virtual object (second frame) and storing these data in locations (e.g., in computer memory) where the data of the first video unit (of the simultaneous frame; first frame) are stored (e.g., in computer memory). This way, less memory is used, and less memory read and write operations need to be carried out. And, upon completion of storing the data representing the representation of the virtual object, the respective frame of the augmented video stream (augmented frame) is “automatically” completed – in the location where initially (and exclusively) the data of the first video stream (first frame) had been stored. The video processing unit can, in this case, merely

read the data from that memory and output the same – as a frame of the augmented video stream (augmented frame).

Nevertheless, this way, the second video stream is factually generated (by the second video unit), as the data representing the representation of the virtual object (or the second frame) are generated, e.g., as a time-sequence of frames, and merely not separately stored, but stored in said locations (computer memory locations) initially taken by data of the first video stream (first frame).

In this regard, generating the second video stream can factually effect the generation of the augmented video stream. And, accordingly, the second video unit can, in part, coincide with (be identical to) the video processing unit, namely in that the generation of the augmented video stream is accomplished by the second video unit.

In any event, it is possible to implement the second video unit and the video processing unit in one and the same unit, such as in one and the same software (program code).

In a straight-forward example for the method, one or more of the following may apply:

- the real object is a smartphone with built-in sensors such as sensors associated with an augmented reality development toolkit, such as with “ARKit” (by Apple) or “ARCore” (by Google) or “Vuforia” (by PTC), as components of the sensing unit;
- the first video unit is a built-in video camera of a computing device (such as of a laptop computer), e.g., built into a monitor of the computing device;
- the computing device embodies (in form of hardware and software) a processing unit of the sensing unit for evaluating raw data (or other sensing data) received from the sensors;
- the real object (such as the smartphone) and the computing device comprise wireless communication capability, such as according to a “Bluetooth” standard or “WiFi”, for transmitting and receiving, respectively, the raw data (or other sensing data) from the sensors;

- the computing device embodies (in form of hardware and software) the second video unit and the video processing unit;
- the augmented video stream can be outputted to the monitor of the computing device and/or can be forwarded to a video conferencing software, such as to be transmitted via the internet or to a peripheral device.

In some embodiments, the method comprises moving the real object in space (i.e. in real space). The moving can be accomplished by a user. In particular, the modulating (modulating the representation of the virtual object in the second video stream) is accomplished in such a way that the representation of the virtual object in the augmented video stream moves in dependence of the movement of the real object. More particularly, the modulating can be accomplished in such a way that the representation of the virtual object in the augmented video stream moves identically to the movement of the real object.

In some way, the real object can in this regard be considered a pointer for the virtual object.

Of course, the determining of the relative pose data takes place during the moving of the real object (since it takes place in real-time).

In some embodiments, the modulating comprises changing the representation of the virtual object in the second video stream in such a way that at least one of

— an apparent position of the virtual object in the augmented video stream is changed in dependence of the relative position in space of the real video camera and the real object;

— an apparent orientation of the virtual object in the augmented video stream is changed in dependence of the relative orientation in space of the real video camera and the real object.

Typically, both applies.

In some embodiments, the modulating comprises changing the representation of the virtual object in the second video stream in such a way that at least one of

— an apparent position of the virtual object in the augmented video stream is linked to the relative position in space of the real video camera and the real object;

— an apparent orientation of the virtual object in the augmented video stream is linked to the relative orientation in space of the real video camera and the real object.

Typically, both applies.

For example, an apparent position of the virtual object in the augmented video stream can change proportionally to changes of the position of the real object, i.e. along the same direction and along a proportional distance.

And/or, for example, an apparent orientation of the virtual object in the augmented video stream can change identically to changes of the orientation of the real object.

In some embodiments, the modulating comprises changing the representation of the virtual object in the second video stream in such a way that changes of the pose of the virtual object in the augmented video stream are identical to changes of the pose of the real object. This way, the movements of the virtual object in the augmented video stream and the movements of the real object (in real space) are coordinated such that the two seem to “move together”. Provided that the real object is represented (visible) in the augmented video stream, this can, in the augmented video stream, provide the illusion that the virtual object is firmly connected to the real object.

In some embodiments, virtual movements of the representation of the virtual object in the augmented video stream can (effectively) be controlled by moving the real object.

In some embodiments, the real object (effectively) functions as a movement controller for controlling virtual movements of the representation of the virtual object in the augmented video stream.

Movements and virtual movements can comprise one or both of position changes and orientation changes, i.e. can comprise pose changes.

In some embodiments, the method comprises

— moving the real object relative to the real video camera.

This can be done, e.g., in particular to make the representation of the virtual object in the augmented video stream move, in particular move in a similar or rather in a corresponding way.

Moving the real object relative to the real video camera can comprise, e.g., moving the real object in real space. And/or during the moving (of the real object relative to the real video camera), the real video camera can remain unmoved (remain still).

In some embodiments, the method further comprises

— positioning the real object in a field-of-view of the real video camera.

This way, the real object may enter a viewport of the real video camera. This can create interesting impressions in the augmented video stream (the real object appearing as a pointer to or holder of the virtual object) and/or be useful for calibration purposes.

In some embodiments, the method further comprises

— carrying out a calibration procedure for facilitating determining the relative pose data.

For example, assuming that pose data of the real object are provided, e.g., determined by the sensing unit, such as by using one or more sensors of the real object, it is likely that said pose data of the real object relate to a coordinate system different from a coordinate system of the real video camera (which is given by the field of view or viewport of the real video camera). Accordingly, it may be necessary (or at least helpful) to determine a transformation which transforms the coordinate system associated with the real object (and the pose data) into the coordinate system associated with the real video camera (or vice versa).

For example, a corresponding calibration procedure can, in this case, comprise:

— positioning the real object distant from the real video camera in a first corner of a viewport of the real video camera and generating and storing first pose data of the real object in this first position;

— positioning the real object distant from the real video camera in a second corner of a viewport of the real video camera opposite the first corner and generating and storing second pose data of the real object in this second position;

— positioning the real object in proximity to the real video and generating and storing third pose data of the real object in this third position;

— determining a transformation from (in dependence of) the first, second and third pose data.

From the three pose data, all three axes of the coordinate system of the viewport can be associated with the coordinate system of the real object.

Typically, the real video camera is assumed to remain unmoved at least during the calibration procedure.

The method for video conferencing can comprise

— generating an augmented video stream as herein described;

— feeding the augmented video stream or a video stream derived therefrom to a video conferencing software.

In some embodiments, the feeding comprises

— feeding the augmented video stream to a device driver software;

the device driver software either

— applying modifications, in particular one or more transformations, to the augmented video stream and feeding the so-modified augmented video stream to the video conferencing software; or

— forwarding the augmented video stream to the video conferencing software.

In particular, the device driver software can be configured such that the only modifications it can apply to a video stream (to the augmented video stream) are transformations. Transformations are format changes, such as changes in color bit depth, changes in the number of pixels per image of the video stream and the like.

The device driver can effect that the augmented video stream is accepted by a computer operating system in the same way as a device driver of a standard (real) video camera (such as the real video camera) is accepted by the computer operating system. This way, the augmented video stream can be made readily available to further computer programs, such as to standard video conferencing software.

The system – comprising the device driver – can be considered to comprise a virtual camera.

In some embodiments, the method further comprises

— the device driver software registering itself with an operating system as a camera device driver.

When the device driver software is configured such that it registers itself with an operating system as a camera device driver upon its installation on a computer on which the operating system is executed, a great simplification for a user is achieved.

Note: When an item is described to be “configured” to carry out a step, this means that concrete measures have been taken which factually enable the item to carry out the step. For example, dedicated program code is implemented enabling the item to carrying out the step when the program code is executed. Thus, this does not include, e.g., the mere

suitability to (possibly) make the item carry out the step, as may be the case for a computer without a dedicated program code.

If not otherwise stated and unless logically impossible, the method steps may be performed in any order (sequence) including simultaneous performance of steps.

As will be readily understood, features mentioned herein with respect to a method can analogously apply for a described apparatus as well, such as for the system. And, vice versa, features mentioned herein with respect to an apparatus (system) can analogously apply for a described method as well. The achievable effects correspond to each other.

The invention comprises apparatuses (systems) with features of corresponding methods according to the invention, and, vice versa, also methods with features of corresponding apparatuses (systems) according to the invention.

The advantages of the apparatuses (systems) basically correspond to the advantages of corresponding methods, and, vice versa, the advantages of the methods basically correspond to the advantages of corresponding apparatuses (systems).

The advantages of the methods basically correspond to the advantages of corresponding apparatuses (systems) and vice versa.

Further embodiments and advantages emerge from the following description and the enclosed figures and from the dependent claims.

Below, the invention is described in more detail by means of examples and the included drawings. In the drawings, same reference numerals refer to same or analogous elements. The figures show schematically:

Fig. 1 a schematic diagram illustrating a system for generating an augmented video stream, also for explication of the corresponding method;

Fig. 2 a schematic illustration of a way of generating an augmented frame;

Fig. 3 a schematic illustration of another way of generating an augmented frame.

The described embodiments are meant as examples or for clarifying the invention and shall not limit the invention.

Fig. 1 shows a schematic diagram of a illustrating a system for generating an augmented video stream S3 which furthermore is used to explain a method for generating the augmented video stream S3.

The system – which can work in real-time – comprises a real object rO, such as a smartphone and a real video camera C which constitutes or is part of a first video unit. Further, it comprises a sensing unit which comprises a processing unit P and one or more sensors (symbolized in Fig. 1 by coordinate systems). The system also comprises a second video unit V and a video processing unit M.

The first video unit generates a first video stream S1, typically showing a real scene visible in a field of view of real video camera C (symbolized in Fig. 1 by thin dotted lines).

The second video unit V generates a second video stream S2 comprising a representation of a virtual object vO.

By video processing unit M, the first video stream S1 and the second video stream S2 are used to create the augmented video stream S3, e.g., by merging the two video streams S1, S2.

The sensing unit (e.g., its processing unit P) outputs relative pose data A which are used by second video unit V for modulating the representation in the second video stream S2 of the virtual object vO.

The relative pose data A characterize a relative position (in real space) and a relative orientation (in real space) of the real object rO and the real video camera C.

If a user now moves the real object rO (the moving symbolized in Fig. 1 by the hollow arrows), the representation of the virtual object vO in the second video stream S2 and in the augmented video stream S3 can change in dependence thereof, e.g., in a corresponding way, as effected by the second video unit V. Thus, the real object rO can be used to control the pose (position and orientation) of the virtual object vO (in the augmented video stream S3).

A way to determine the relative pose data A comprises determining the pose of the real object rO, e.g., by means of sensors comprised in the real object, such as sensors associated with an augmented reality development toolkit, such as "ARKit" in case the real object is an "Apple" "iPhone 12 Pro". Pose data of the real object rO obtained in such a way however are not the sought relative pose data A. However, a link between a coordinate system on which the pose data of the real object rO are based and a coordinate system associated with the real video camera C can make possible to transform the pose data of the real object rO into the relative pose data A. Such a link can be accomplished in a calibration procedure. Provided with the pose data of the real object rO (usually in a wireless fashion, such as via "Bluetooth" or "WiFi") and with the calibration information, processing unit P can determine the relative pose data A. This works well, at least as long as the pose of real video camera C remains unchanged, i.e. as long as real video camera C is still (not moved).

An alterantive way of determining the relative pose data A is to determine the pose data of the real object rO and the pose data of the real video camera C, as symbolized in Fig. 1, both with respect to one and the same coordinate system. Therefrom, the relative pose data A can be readily determined, possibly even without requiring a calibration procedure. This way works well also when real video camera C is moved.

Various further ways of sensing to finally derive the relative pose data A are enabled, in particular considering today's commercially available or built-in sensors, which comprise, e.g., depth sensing techniques, like based on time-of-flight or structured light or stereo vision or monocular vision in combination with algorithmic methods, machine

learning-based methods, deep learning-based methods, artificial intelligence-based methods. Also the first video stream S1 may be used as a sensor of the sensing unit, e.g., the sensing being based on image analysis with object recognition and/or algorithmic methods, machine learning-based methods, deep learning-based methods, artificial intelligence-based methods such as to at least partially determine the relative pose data – which bears the advantage that a so-detected pose of the real object rO already can be – intrinsically – the relative pose (because the images, forming the basis for the determination of the relative pose data are always in the coordinate system of the real video camera). When the real video camera C is a 2D-color-plus-depth camera or even a 3D video camera, the relative pose data A may be even more precisely determinable.

The processing unit P, the second video unit V and the video processing unit M may, for example, be implemented in software (program code) implemented in a computer of the system, such as on a laptop or desktop computer or a mobile, e.g., handheld or head-mounted computing device, symbolized in Fig. 1 by the large dashed rectangle.

Another piece of software (program code) may be comprised in the system and may be implemented in the computer, functioning as a device driver to be recognized by an operating system running in the computer as a camera device driver. In Fig. 1, the device driver is symbolized as a small dashed rectangle. It receives the augmented video stream S3 and outputs it (or a video stream derived from augmented video stream S3, e.g., by applying format changes), so that it can be readily fed to further programs such as to a standard video conferencing software.

The real video camera C can be, e.g., a camera of the computer, such as a built-in camera of a monitor of a desktop computer or the built-in camera of a laptop computer. Or it can be a camera operationally connected to the computer, such as a peripheral camera device, e.g., connected to the computer in a wirebound or wireless fashion.

Figs. 2 and 3 each show a schematic illustration of a way of generating an augmented frame F3, i.e. a frame of the augmented video stream S3. A first frame, i.e. a frame of

the first video stream S1, is labelled F1, and a second frame, i.e. a frame of the second video stream S2, is labelled F2. In first frame F1, a representation of real-time video stream data from real video camera C is illustrated. In second frame F2, a representation of the virtual object vO is illustrated. In augmented frame F3, a result of merging frames F1 and F2 is illustrated.

Fig. 2 is to illustrate that frame F1 is stored in a first memory location and that frame F2 is stored in a second memory location. Then, by video processing unit M, frame F3 is generated in a third memory location, from frames F1 and F2.

Fig. 3 is to illustrate that frame F1 is stored in a first memory location. Then, during generation of the data representing the representation of the virtual object vO, these data are written (stored) into said first memory location, e.g., by simply overwriting the corresponding data of the first frame F1; the middle portion of Fig. 3 illustrates the situation after a bit more than half of said data representing the representation of the virtual object vO have been generated and stored in the first memory location. And finally (cf. right portion of Fig. 3), all data representing the representation of the virtual object vO are generated and stored in said first memory location – and thus, the augmented frame F3 is generated from frames F1 and F2.

In a not specifically shown alternative way of proceeding, one starts like in the procedure of Fig. 2, i.e. with frame F1 stored in a first memory location, and frame F2 is stored in a second memory location. Then, however, the (not fully transparent) pixels (or voxels) of frame F2 are stored (similarly to the procedure of Fig. 3) in the first memory location F1 or in the second memory location F2.

(A possible handling of merely partially transparent pixels has been described further above.)

The invention makes possible intuitive and simple-to-use virtual object modifications and uses of the generated augmented video stream.

Aspects of the embodiments have been described in terms of functional units. As is readily understood, these functional units may generally be realized in virtually any number of hardware and/or software components adapted to performing the specified functions.

5

Patent Claims:

1. Method for generating an augmented video stream (S3), comprising

— providing a real video camera (C);

— generating a first video stream (S1) comprising real-time video stream data from the real video camera (C);

— providing a real object (rO);

— determining in real-time relative pose data (A) indicative of a relative position in space and of a relative orientation in space of the real video camera (C) and the real object (rO);

— generating a second video stream (S2) comprising a representation of a virtual object (vO);

— modulating the representation of the virtual object (vO) in the second video stream (S2) in real-time and in dependence of the relative pose data (A);

— outputting and/or generating from the first video stream (S1) and the second video stream (S2) the augmented video stream (S3) in real-time.

2. The method according to claim 1, comprising moving the real object (rO) in space, in particular wherein the modulating is accomplished in such a way that the representation of the virtual object (vO) in the augmented video stream (S3) moves in dependence of the movement of the real object (rO), more particularly moves identically to the movement of the real object (rO).

3. The method according to claim 1 or claim 2, wherein the modulating comprises changing the representation of the virtual object (vO) in the second video stream (S2) in such a way that at least one of

— an apparent position of the virtual object (rO) in the augmented video stream (S3) is changed in dependence of the relative position in space of the real video camera (C) and the real object (rO);

— an apparent orientation of the virtual object (rO) in the augmented video stream (S3) is changed in dependence of the relative orientation in space of the real video camera (C) and the real object (rO).

4. The method according to one of claims 1 to 3, wherein the modulating comprises changing the representation of the virtual object (vO) in the second video stream (S2) in such a way that at least one of

— an apparent position of the virtual object (rO) in the augmented video stream (S3) is linked to the relative position in space of the real video camera (C) and the real object (rO);

— an apparent orientation of the virtual object (rO) in the augmented video stream (S3) is linked to the relative orientation in space of the real video camera (C) and the real object (rO).

5. The method according to one of claims 1 to 4, further comprising

— moving the real object (rO) relative to the real video camera (C), in particular to make the representation of the virtual object (vO) in the augmented video stream (S3) move in a corresponding way.

6. The method according to one of claims 1 to 5, further comprising

— positioning the real object in a field-of-view of the real video camera.

7. The method according to one of claims 1 to 6, further comprising

— carrying out a calibration procedure for facilitating determining the relative pose data.

8. The method according to one of claims 1 to 7, wherein determining the pose data comprises

— real-time sensing position in space and/or orientation in space of the real object;

— generating sensing data indicative of results of the sensing.

9. The method according to claim 8, wherein the sensing takes place in the real object (rO), in particular wherein the sensing is accomplished by one or more sensors comprised in the real object (rO);

the method further comprising

— wirelessly transmitting the sensing data;

in particular wherein the sensing data are transmitted to a processing unit (P) for determining in real-time the pose data (A) in dependence of the sensing data, more particularly wherein the processing unit (P) is comprised in a computing device, and the real video camera is a built-in camera of the computing device or a camera operationally connected to the computing device.

10. The method according to one of claims 1 to 9, comprising

— controlling virtual movements of the representation of the virtual object (vO) in the augmented video stream (S3) by moving the real object (rO).

11. Method for video conferencing, comprising

— generating an augmented video stream (S3) according to one of the preceding claims;

— feeding the augmented video stream (S3) or a video stream derived therefrom to a video conferencing software.

12. The method according to claim 11, wherein the feeding comprises

— feeding the augmented video stream (S3) to a device driver software;

the device driver software either

— modifying, in particular transforming, the augmented video stream (S3) and feeding the modified augmented video stream to the video conferencing software; or

— forwarding the augmented video stream (S3) to the video conferencing software.

13. The method according to claim 12, further comprising:

— the device driver software registering itself with an operating system as a camera device driver.

14. System for generating an augmented video stream (S3), comprising

— a first video unit comprising a real video camera (C), configured to generate a first video stream (S1) comprising real-time video stream data from the real video camera (C);

— a real object (rO);

— a sensing unit configured to determine in real-time relative pose data (A) indicative of a relative position in space and of a relative orientation in space of the real video camera (C) and the real object (rO);

— a second video unit (V) operationally connected to the sensing unit for receiving the relative pose data (A) from the sensing unit and configured

- to generate a second video stream (S2) comprising a representation of a virtual object (vO); and
- to modulate the representation of the virtual object (vO) in the second video stream (S2) in real-time and in dependence of the relative pose data (A);

— a video processing unit (M) operationally connected to the first video unit and to the second video unit (V), configured to output and/or generate from the first video stream (S1) and the second video stream (S2) the augmented video stream (S3) in real-time.

15. The system according to claim 14, wherein the sensing unit comprises one or more sensors for position sensing and/or orientation sensing, and in particular wherein said one or more sensors are comprised in the real object (rO).

16. The system according to claim 14 or 15, wherein the real object is a mobile computing device, in particular a smartphone or a tablet computer.

17. The system according to one of claims 14 to 15, wherein the real object is an add-on device for mobile computing device, in particular wherein the mobile computing device is a smartphone or a tablet computer.

18. The system according to one of claims 14 to 17, comprising a computing device, in particular a laptop computer or a desktop computer, wherein the second video unit (V) and the video processing unit (M) are comprised in the computing device, in particular in form of computer program code, which when executed in the computing device performs the respective steps.

19. The system according to claim 18, wherein the real video camera (C) is a component of the computing device, in particular wherein the first video unit is component of the computing device.

20. The system according to claim 18 or 19, wherein the real video camera (C) is a peripheral camera operationally connectable the computing device.

21. The system according to one of claims 18 to 20, wherein at least a portion of the sensing unit, in particular a processing unit (P) of the sensing unit, is comprised in the computing device.

22. The system according to claim 15 AND one of claims 18 to 21, the sensing unit comprising a processing unit (P) which is comprised in the computing device, wherein the processing unit (P) is configured to process sensing data indicative of results of the sensing of the one or more sensors, in particular wherein the sensing unit is implemented in the computing device in form of computer program code, which when executed in the computing device performs the respective steps.

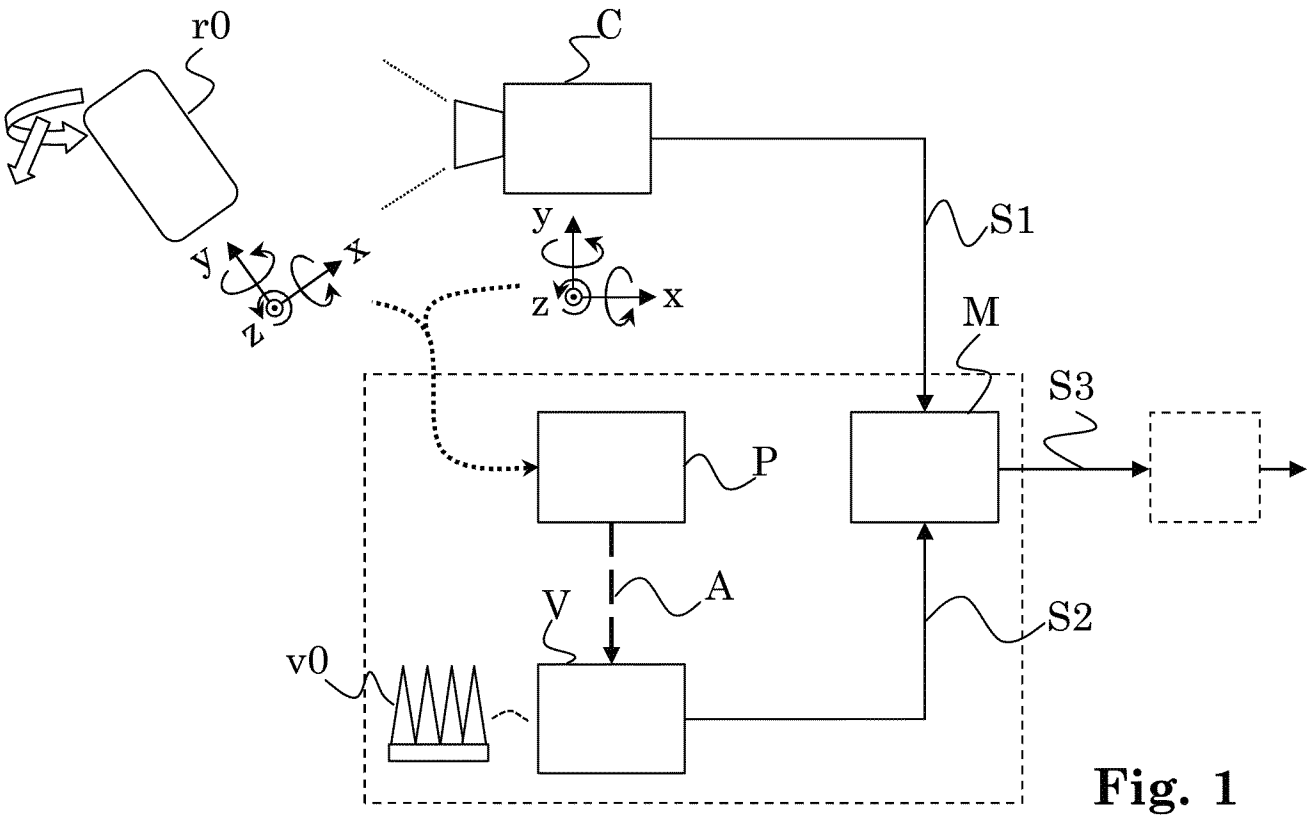

Fig. 1

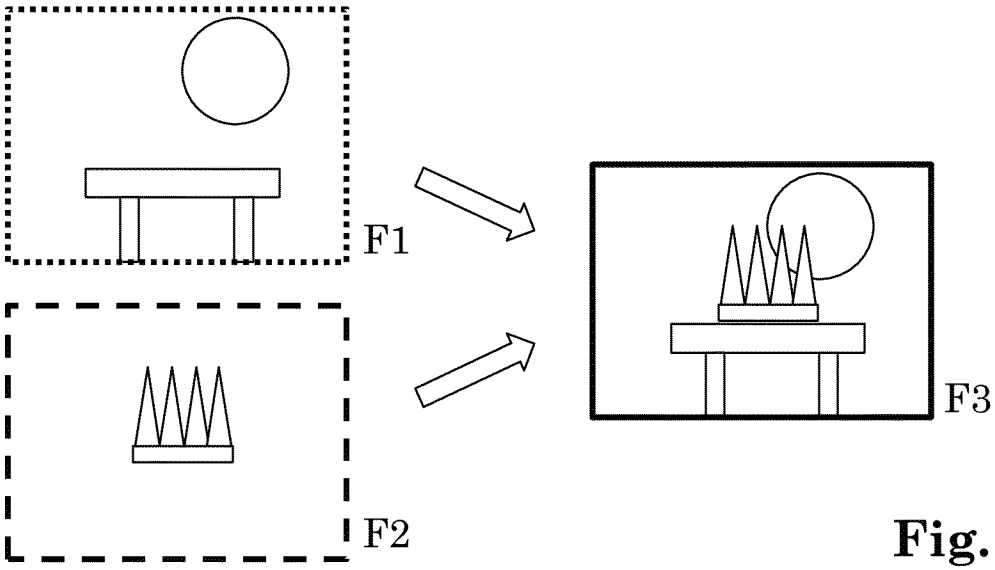

Fig. 2

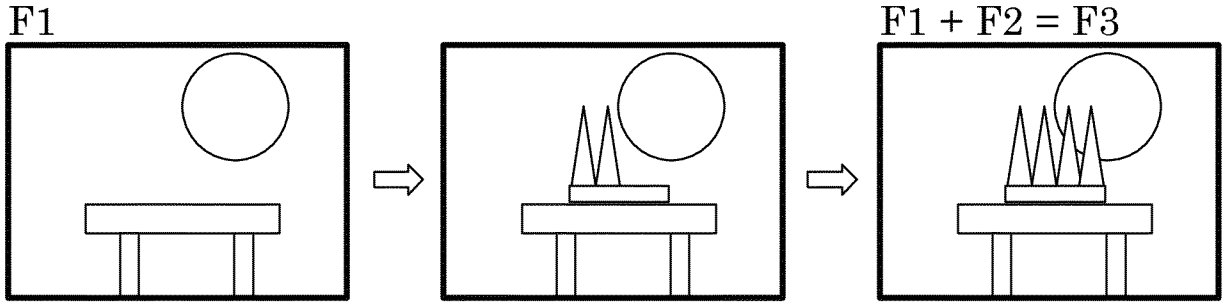

Fig. 3

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2022/070186

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06T19/00
ADD.

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06T H04N

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

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	CN 111 970 535 A (XMOV SHANGHAI INFORMATION TECH CO LTD; SHANGHAI MOWU TECH CO LTD) 20 November 2020 (2020-11-20)	1-15, 18-22
Y	paragraph [0183] - paragraph [0184]; figure 5	16,17

X	Zhanpeng Huang ET AL: "Mobile augmented reality survey: a bottom-up approach", , 17 September 2013 (2013-09-17), XP055134545, Retrieved from the Internet: URL:http://arxiv.org/abs/1309.4413 [retrieved on 2019-07-26]	1-15, 18-22
Y	Paragraphs 2.3.2, 2.4.1 and 2.4.2	16,17

	-/--	

[x] Further documents are listed in the continuation of Box C.

[x] See patent family annex.

* Special categories of cited documents :

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

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

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

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

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

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

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

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

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
14 October 2022	24/10/2022

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

1

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2022/070186

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	ADRIAN DAVID CHEOK ET AL: "Human Pacman: a mobile, wide-area entertainment system based on physical, social, and ubiquitous computing", PERSONAL AND UBIQUITOUS COMPUTING, SPRINGER VERLAG, LONDON, GB, vol. 8, no. 2, 1 May 2004 (2004-05-01), pages 71-81, XP058126351, ISSN: 1617-4909, DOI: 10.1007/S00779-004-0267-X	1-15, 18-22
Y	Paragraph 3 "System design and game play"	16,17
Y	ANDERS HENRYSSON ET AL: "Mobile phone based AR scene assembly", MOBILE AND UBIQUITOUS MULTIMEDIA, ACM, 2 PENN PLAZA, SUITE 701 NEW YORK NY 10121-0701 USA, 8 December 2005 (2005-12-08), pages 95-102, XP058363160, DOI: 10.1145/1149488.1149504 ISBN: 978-0-473-10658-4 Table 1, page 98	16,17

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2022/070186

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
CN 111970535 A	20-11-2020	CN 111970535 A	20-11-2020
		WO 2022062678 A1	31-03-2022