US 20240212310A1

(19) United States

(12) Patent Application Publication

DOMANSKI et al.

(10) Pub. No.: US 2024/0212310 A1

(43) Pub. Date: Jun. 27, 2024

(54) METHOD AND SYSTEM FOR PROCESSING VIDEO

(71) Applicant: **GUANGDONG OPPO MOBILE TELECOMMUNICATIONS CORP., LTD.**, Dongguan (CN)

(72) Inventors: **Marek DOMANSKI**, Poznan (PL); **Tomasz GRAJEK**, Poznan (PL); **Adam GRZELKA**, Poznan (PL); **Slawomir MACKOWIAK**, Poznan (PL); **Slawomir ROZEK**, Poznan (PL); **Olgierd STANKIEWICZ**, Poznan (PL); **Jakub STANKOWSKI**, Poznan (PL)

(21) Appl. No.: **18/600,755**

(22) Filed: **Mar. 10, 2024**

Related U.S. Application Data

(63) Continuation of application No. PCT/CN2022/077142, filed on Feb. 21, 2022.

(30) **Foreign Application Priority Data**

Sep. 13, 2021 (EP) 21461589.0

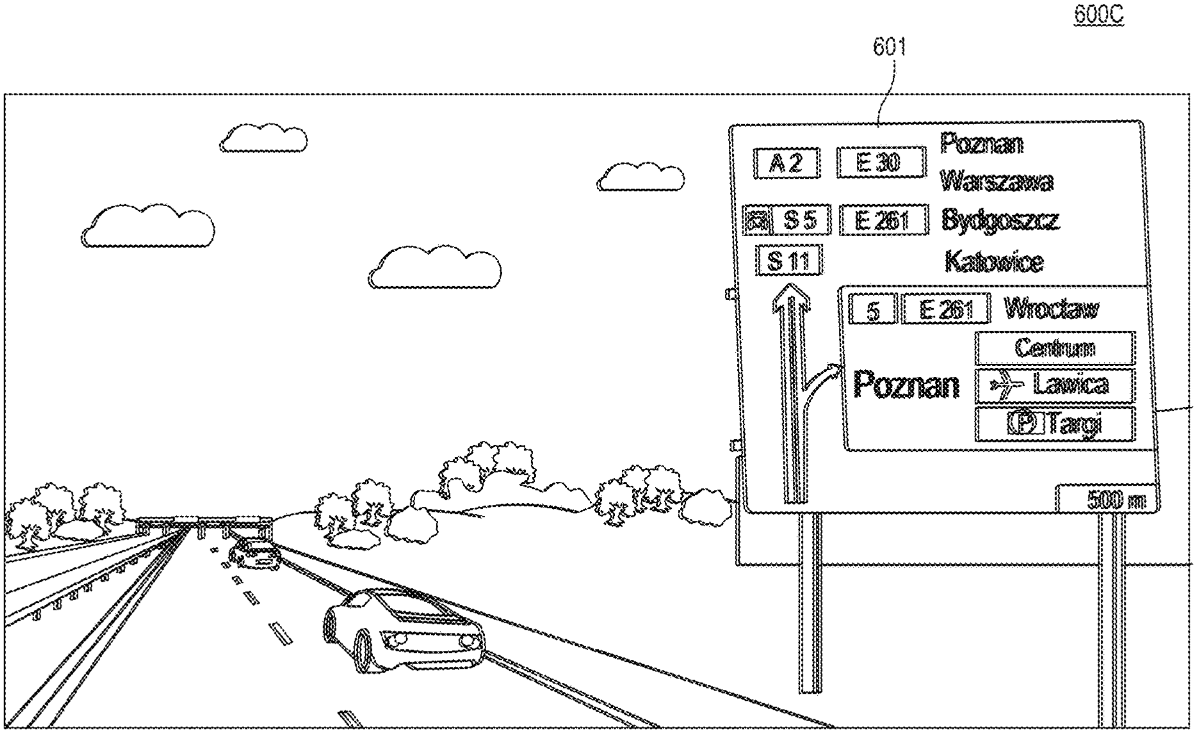
Publication Classification

(51) **Int. Cl.**
G06V 10/44 (2006.01)

(52) **U.S. Cl.**
CPC *G06V 10/44* (2022.01); *G06V 2201/07* (2022.01)

(57) ABSTRACT

Systems and methods for processing a video are disclosed herein. The method include (1) receiving an input video; (2) identifying one or more objects in the input video; (3) determining a set of feature descriptors associated with the one or more objects; (4) generating a set of feature units associated with the one or more objects.



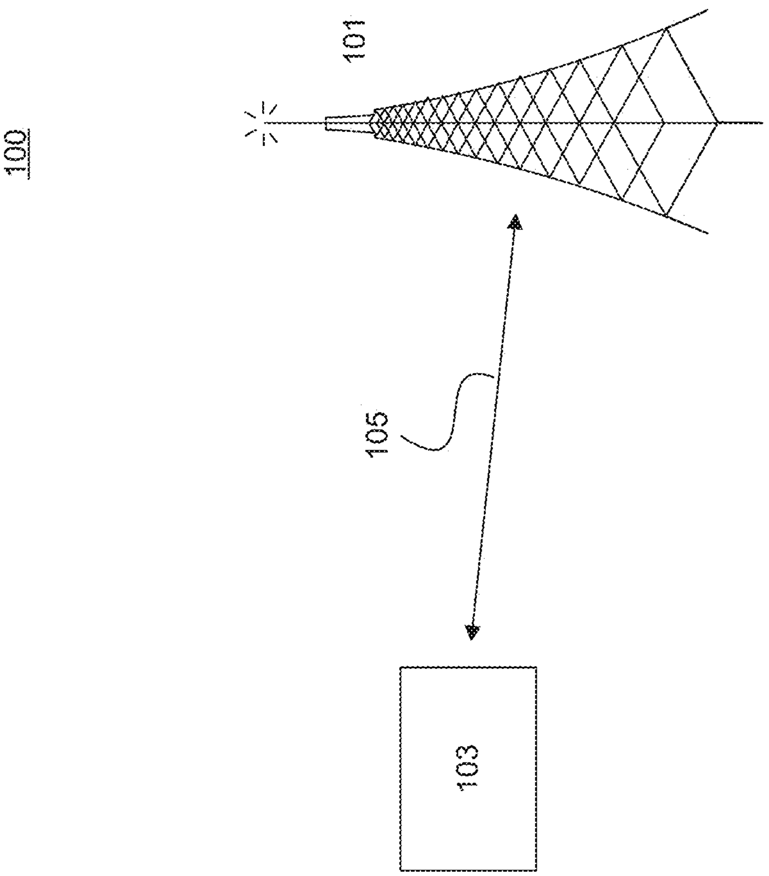


FIG. 1

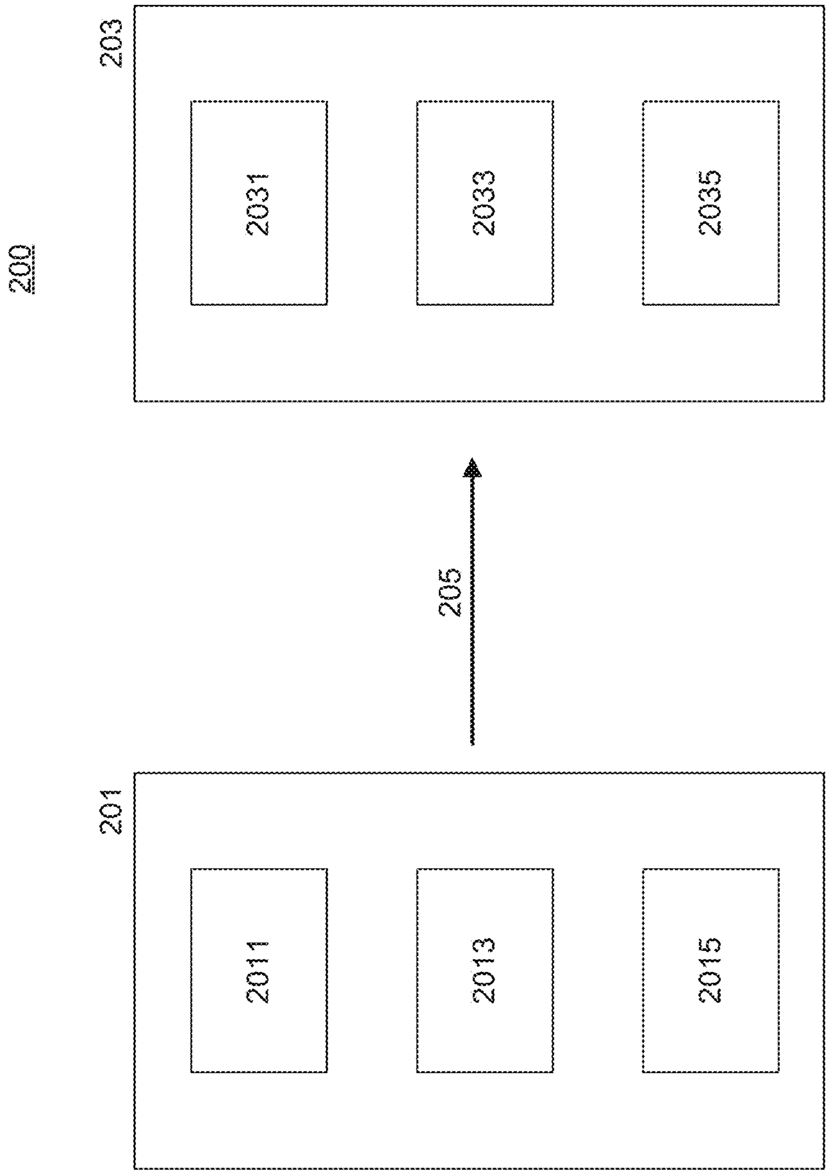


FIG. 2A

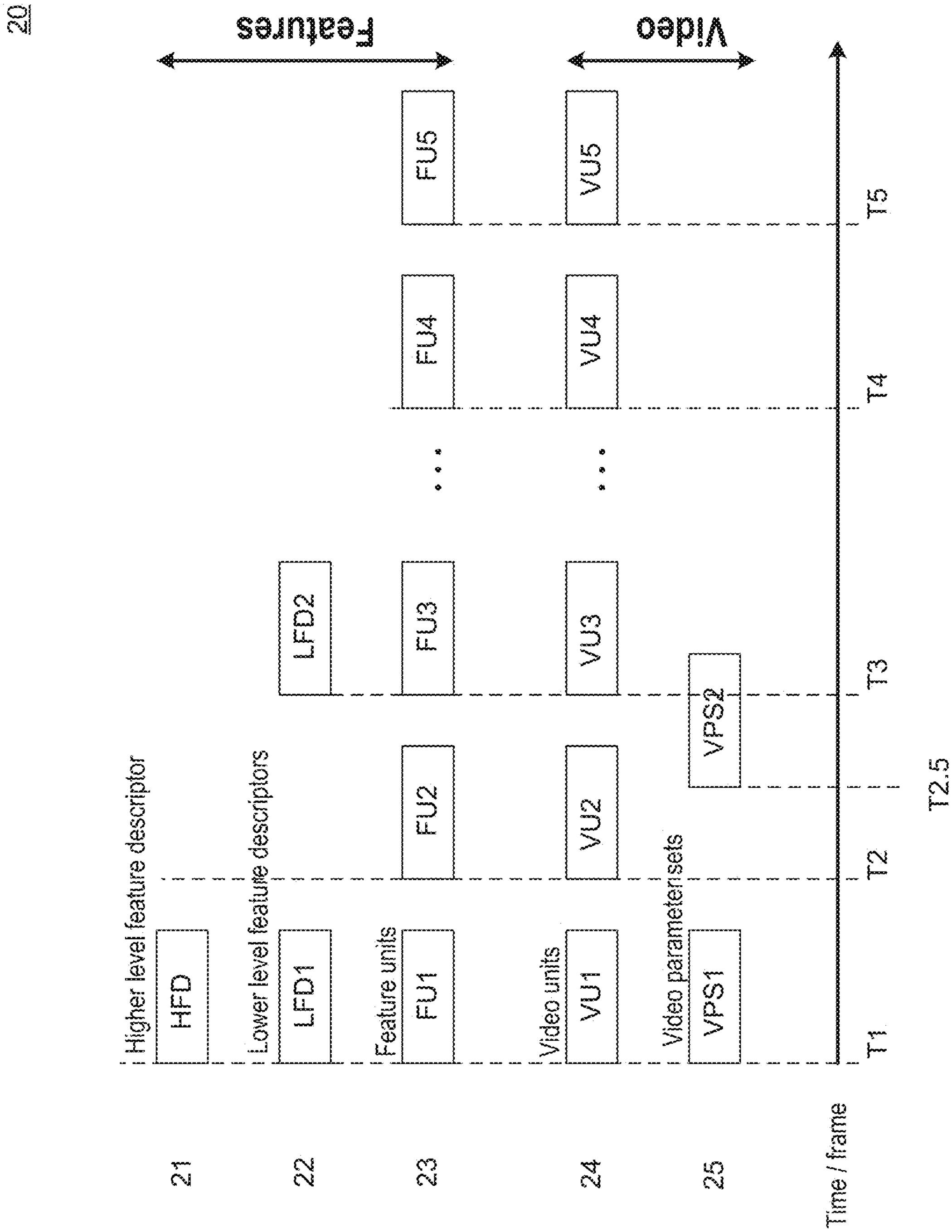

20
21
22
23
24
25
Time / frame
Higher level feature descriptor
HFD
Lower level feature descriptors
LFD1
LFD2
Feature units
FU1
FU2
FU3
FU4
FU5
Video units
VU1
VU2
VU3
VU4
VU5
Video parameter sets
VPS1
VPS2
T1
T2
T2.5
T3
T4
T5
Features
Video

FIG. 2B

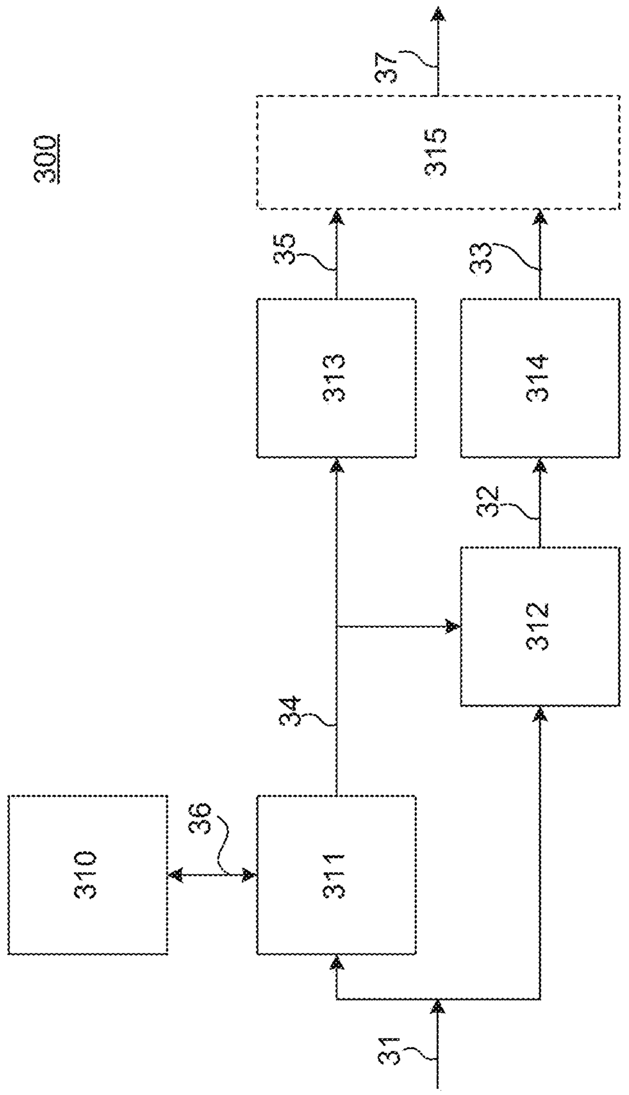


FIG. 3

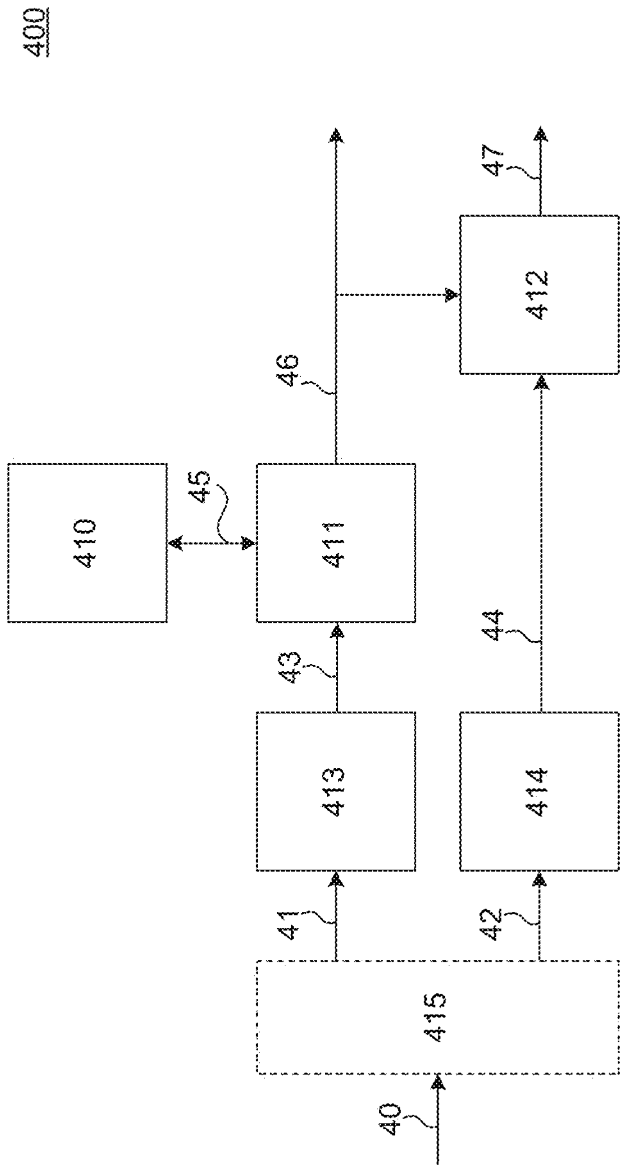


FIG. 4

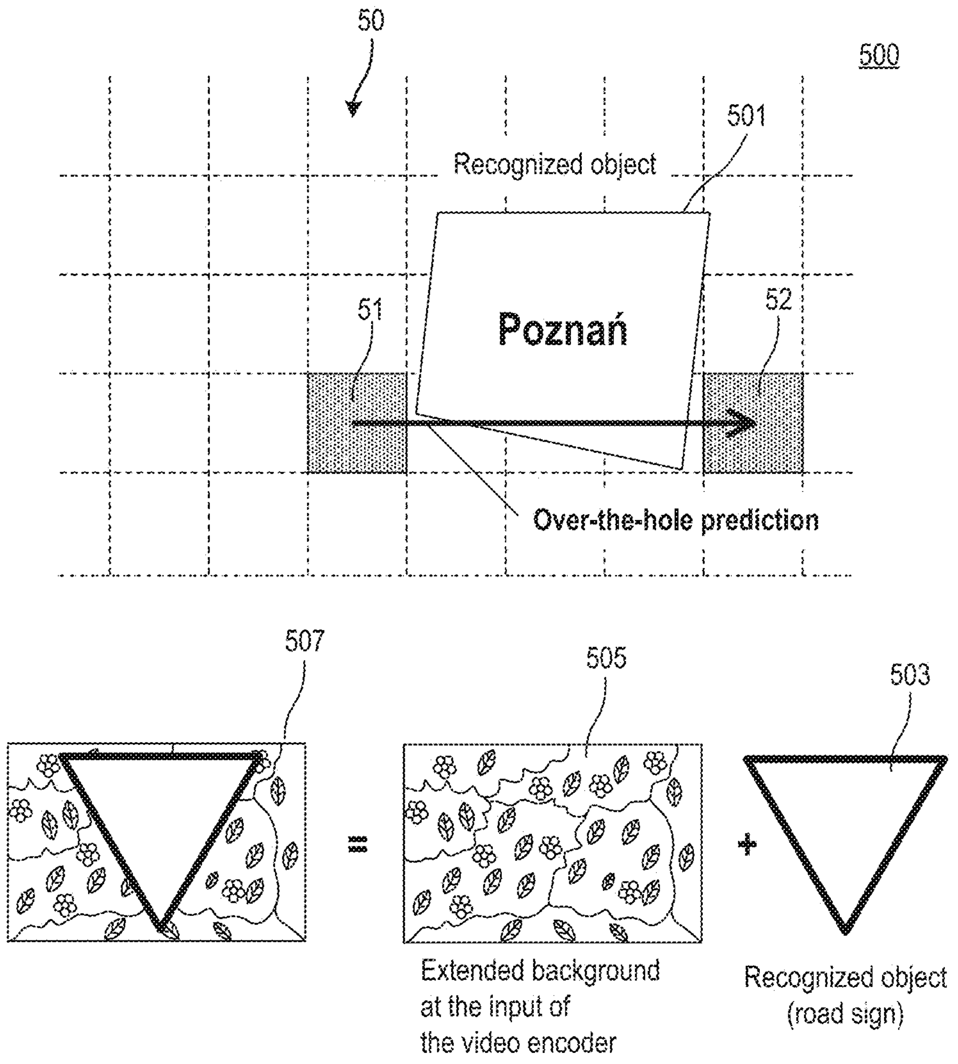

50
500
501
Recognized object
Poznań
51
52
Over-the-hole prediction
507
505
503
=
+
Extended background
at the input of
the video encoder
Recognized object
(road sign)

FIG. 5

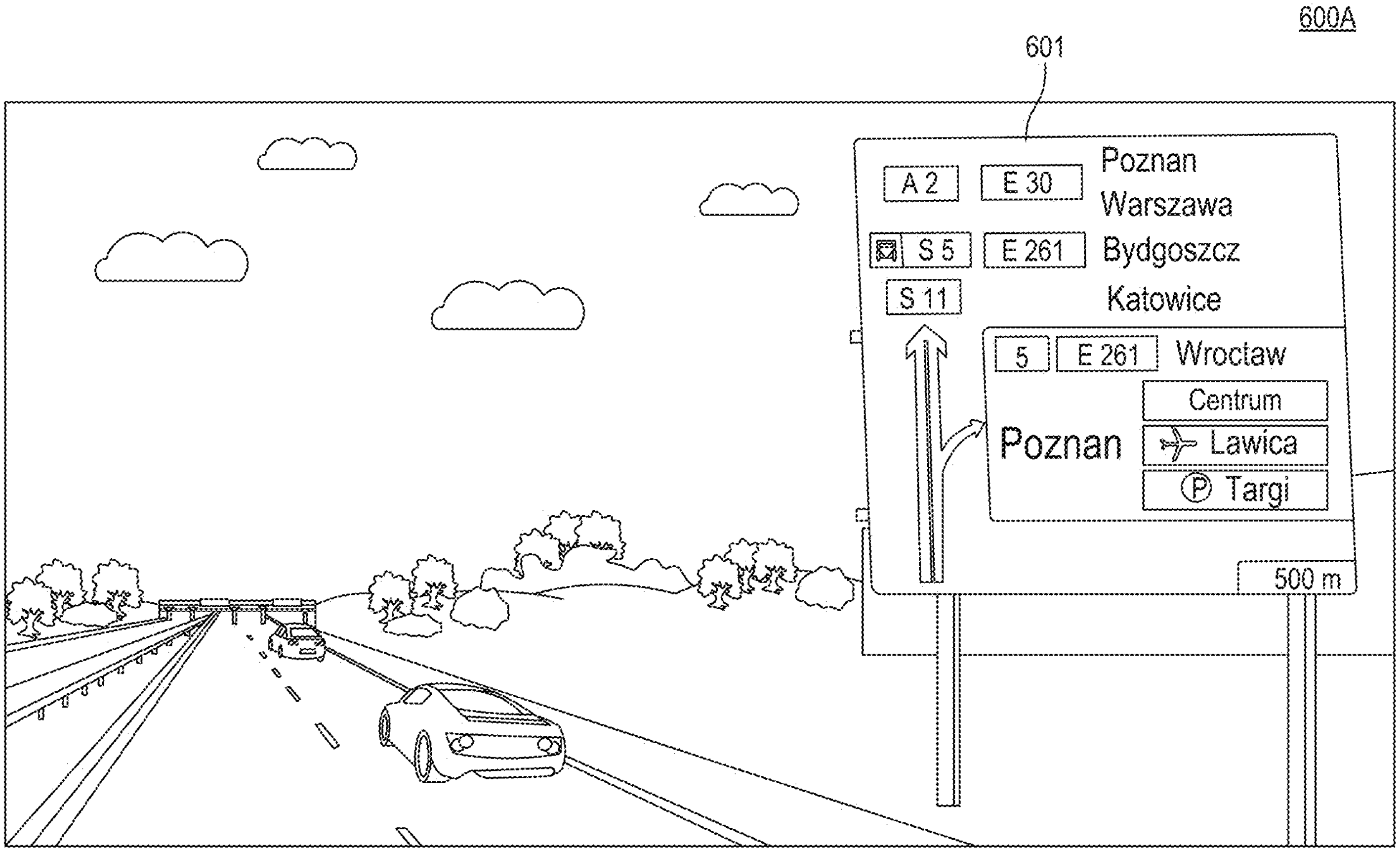

600A
601
A 2
E 30
Poznan
Warszawa
S 5
E 261
Bydgoszcz
S 11
Katowice
5
E 261
Wrocław
Centrum
Poznan
Lawica
Ⓟ Targi
500 m

FIG. 6A

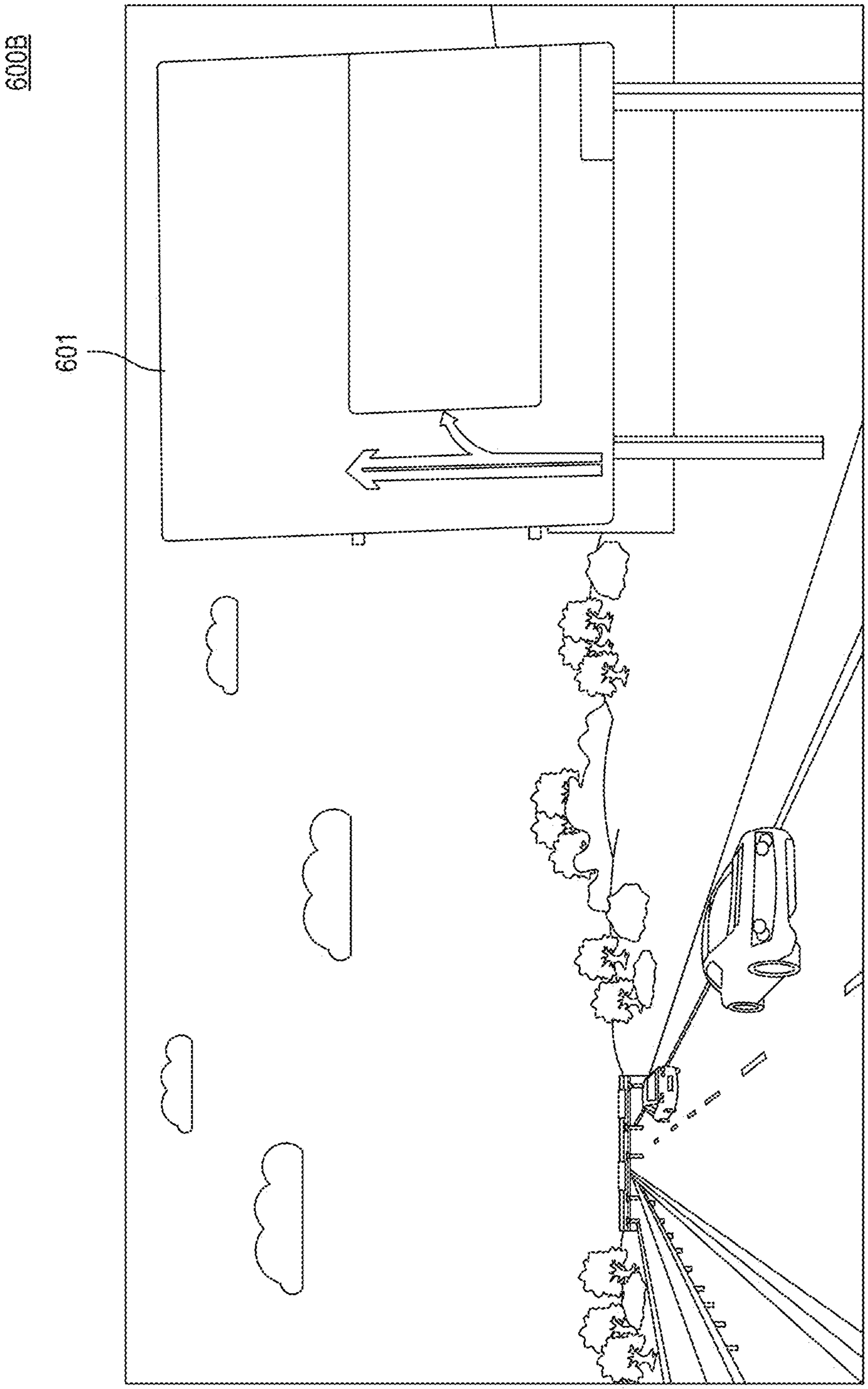


FIG. 6B

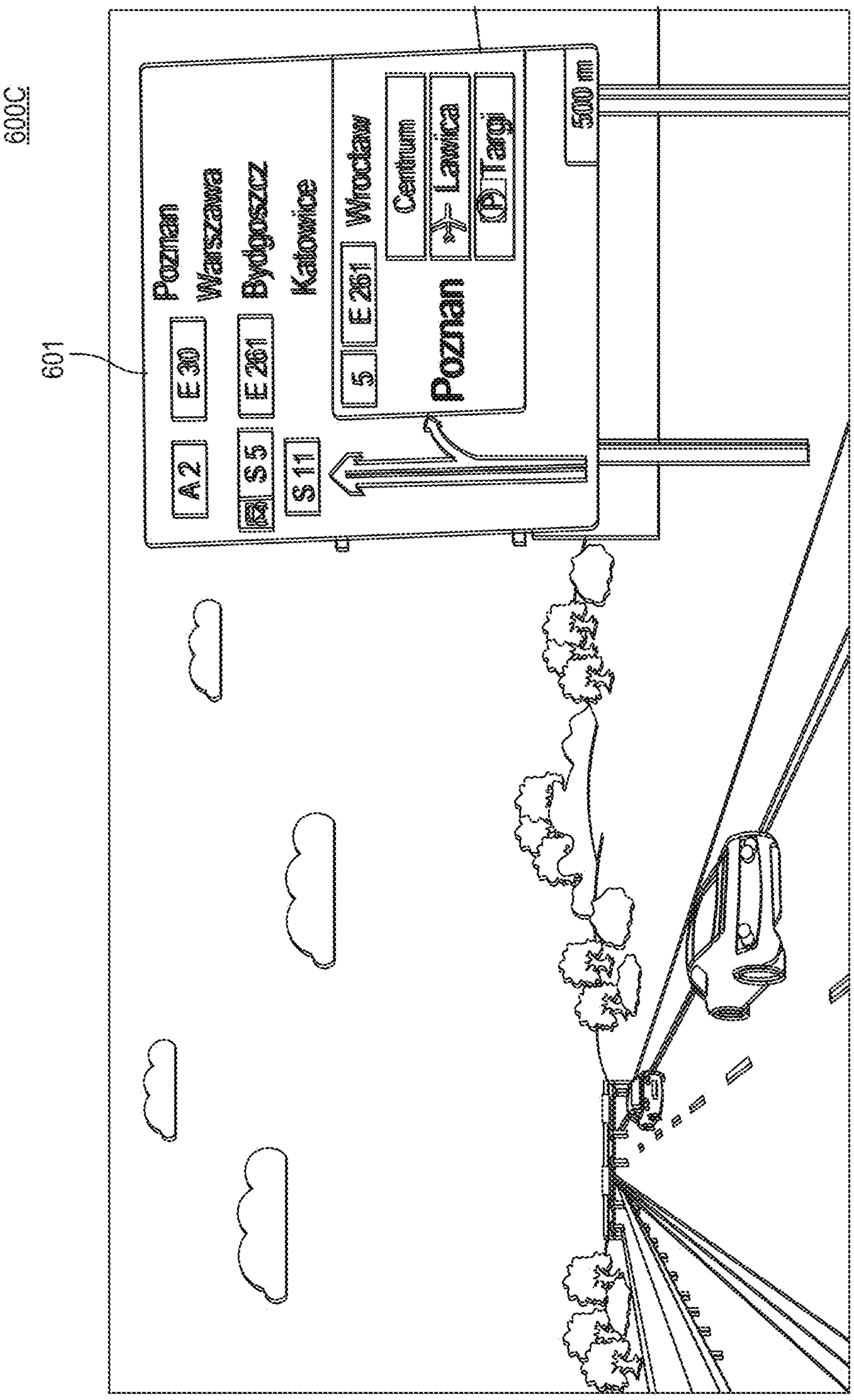

600C
601
A 2
E 30
Poznan
Warszawa
S 5
E 261
Bydgoszcz
S 11
Katowice
5
E 261
Wroclaw
Poznan
Centrum
Lawica
Targi
500 m

FIG. 6C

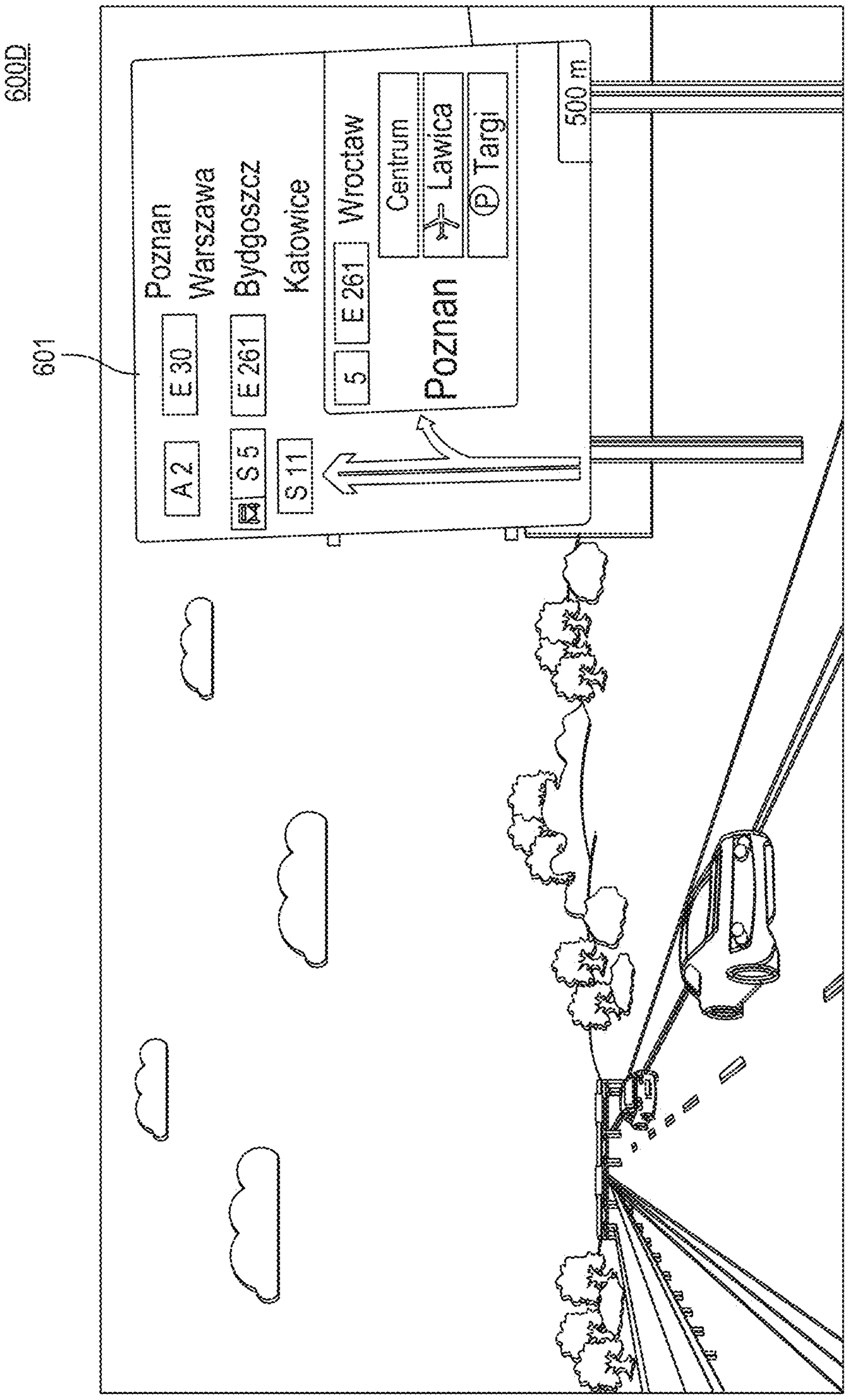

600D
601
A 2
E 30
Poznan
Warszawa
S 5
E 261
Bydgoszcz
Katowice
S 11
5
E 261
Wroctaw
Poznan
Centrum
Lawica
P Targi
500 m

FIG. 6D

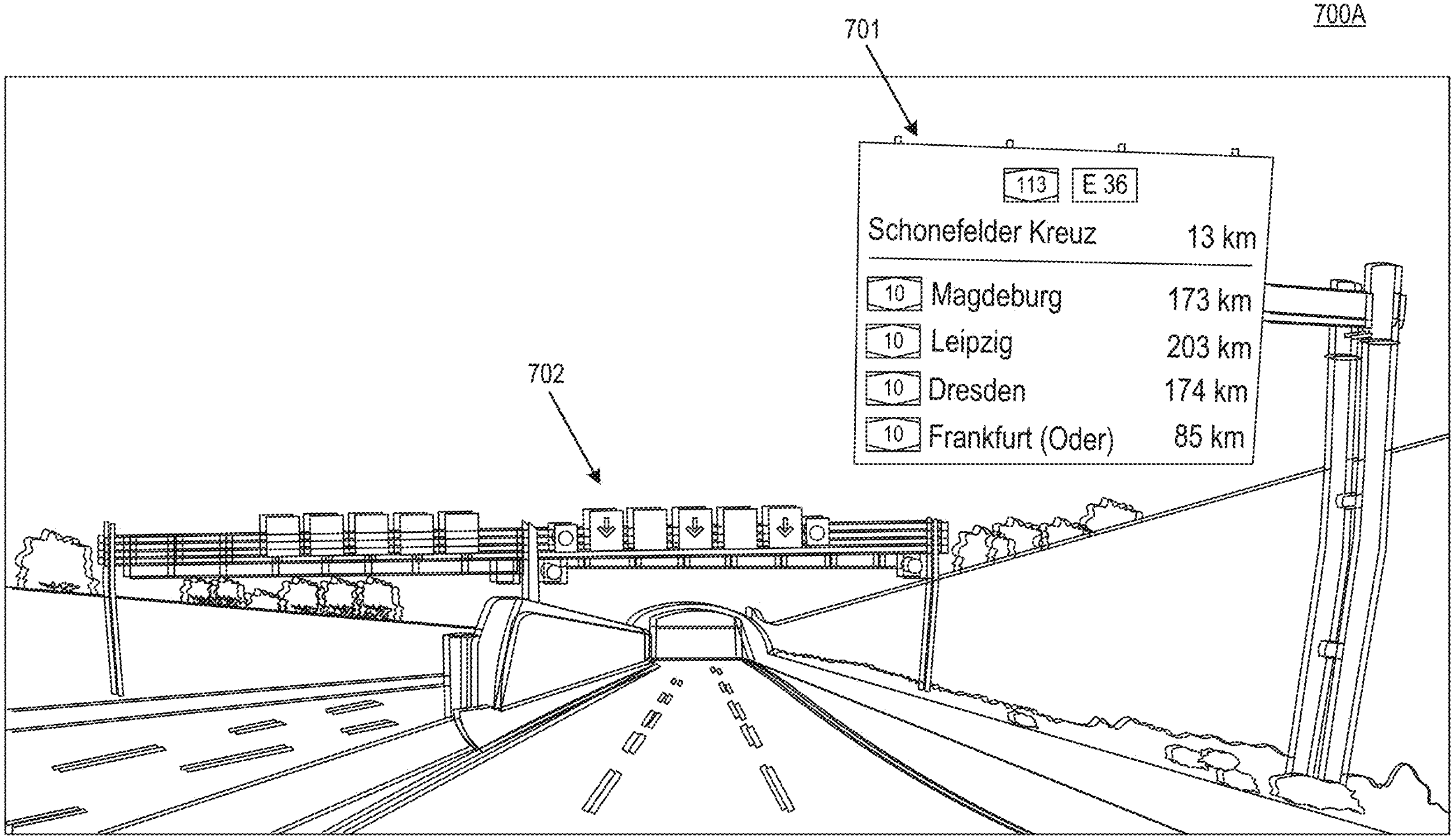

700A
701
702
113
E 36
Schonefelder Kreuz 13 km
10 Magdeburg 173 km
10 Leipzig 203 km
10 Dresden 174 km
10 Frankfurt (Oder) 85 km

FIG. 7A

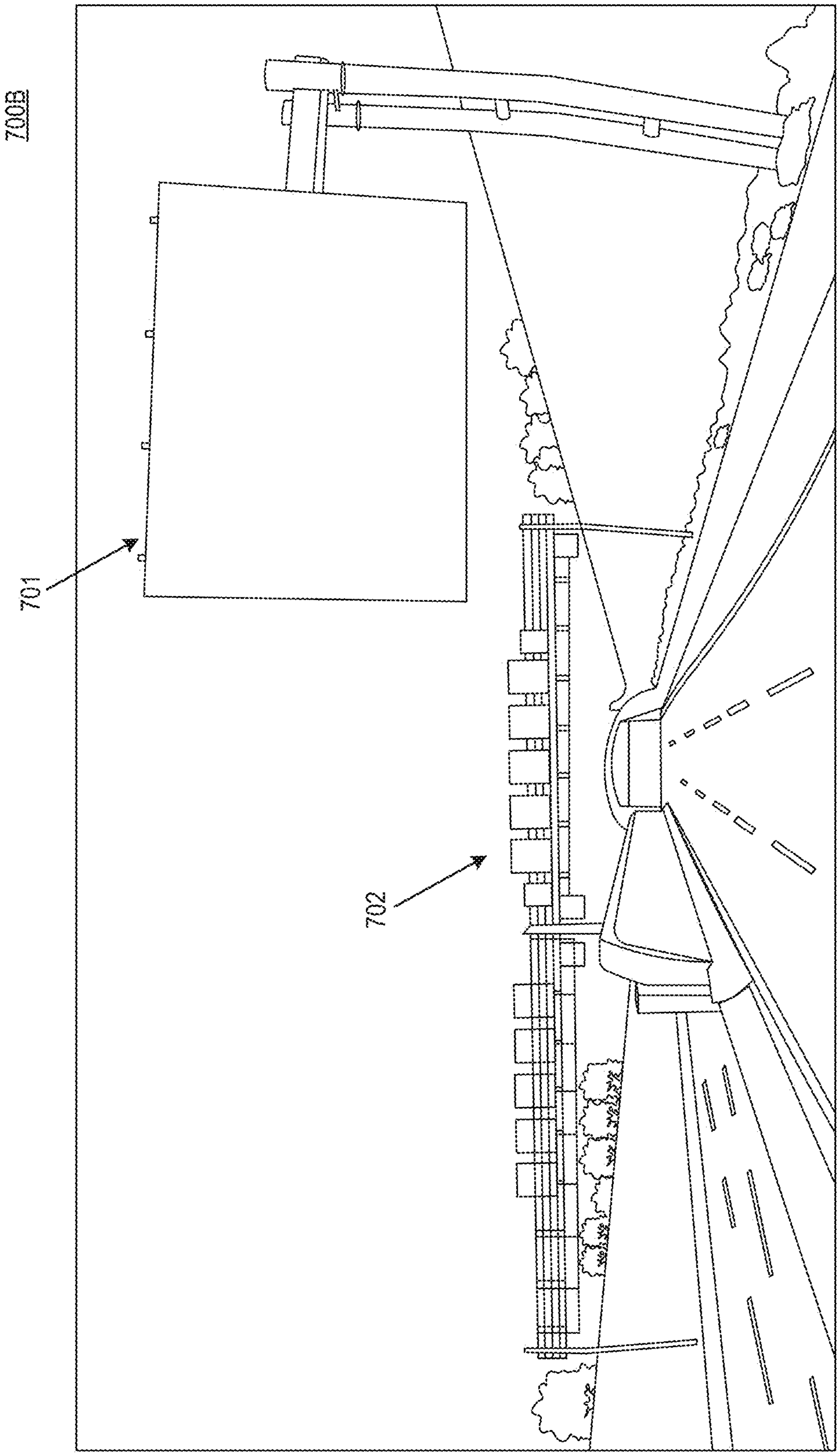


FIG. 7B

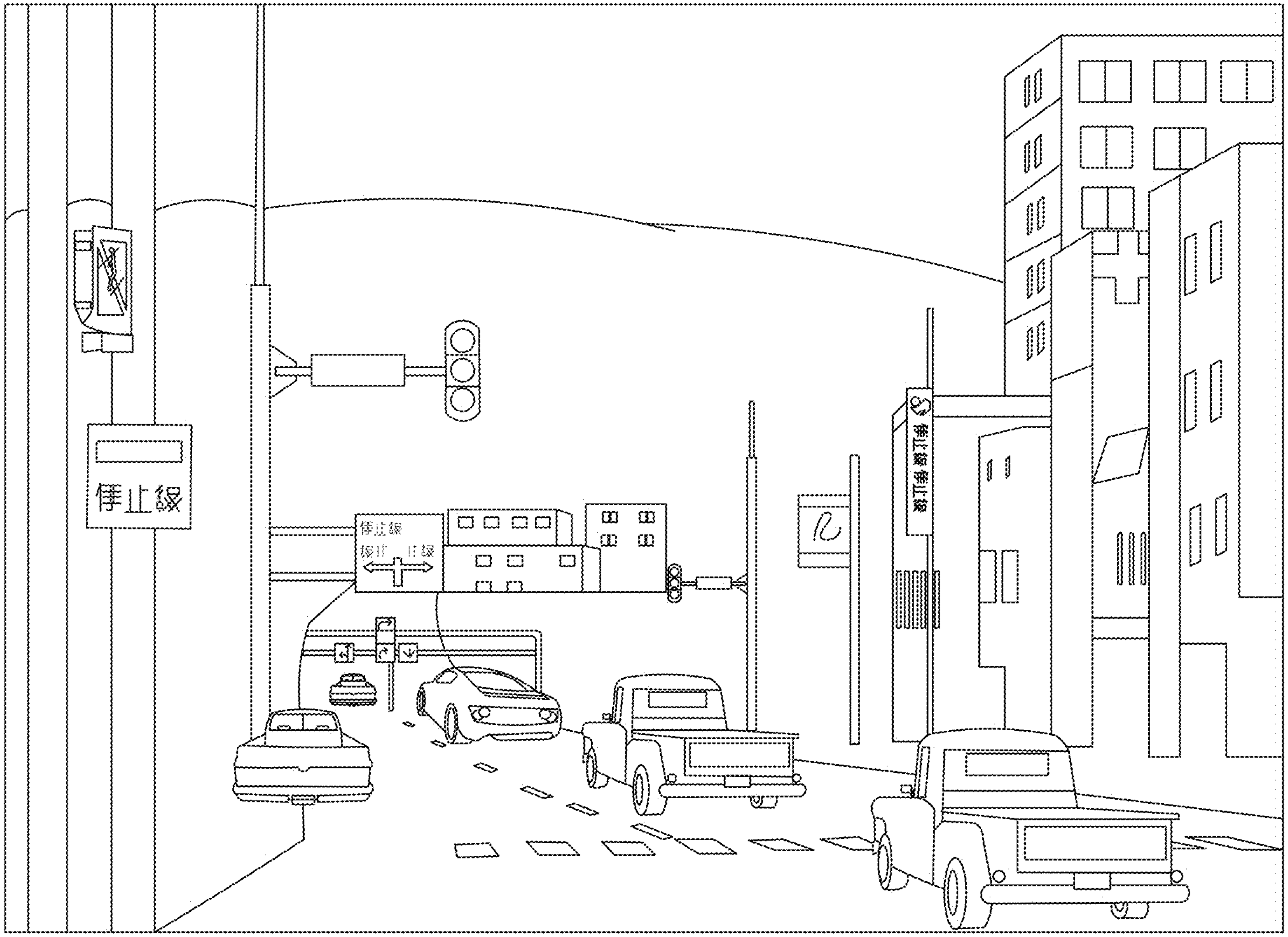


FIG. 8A

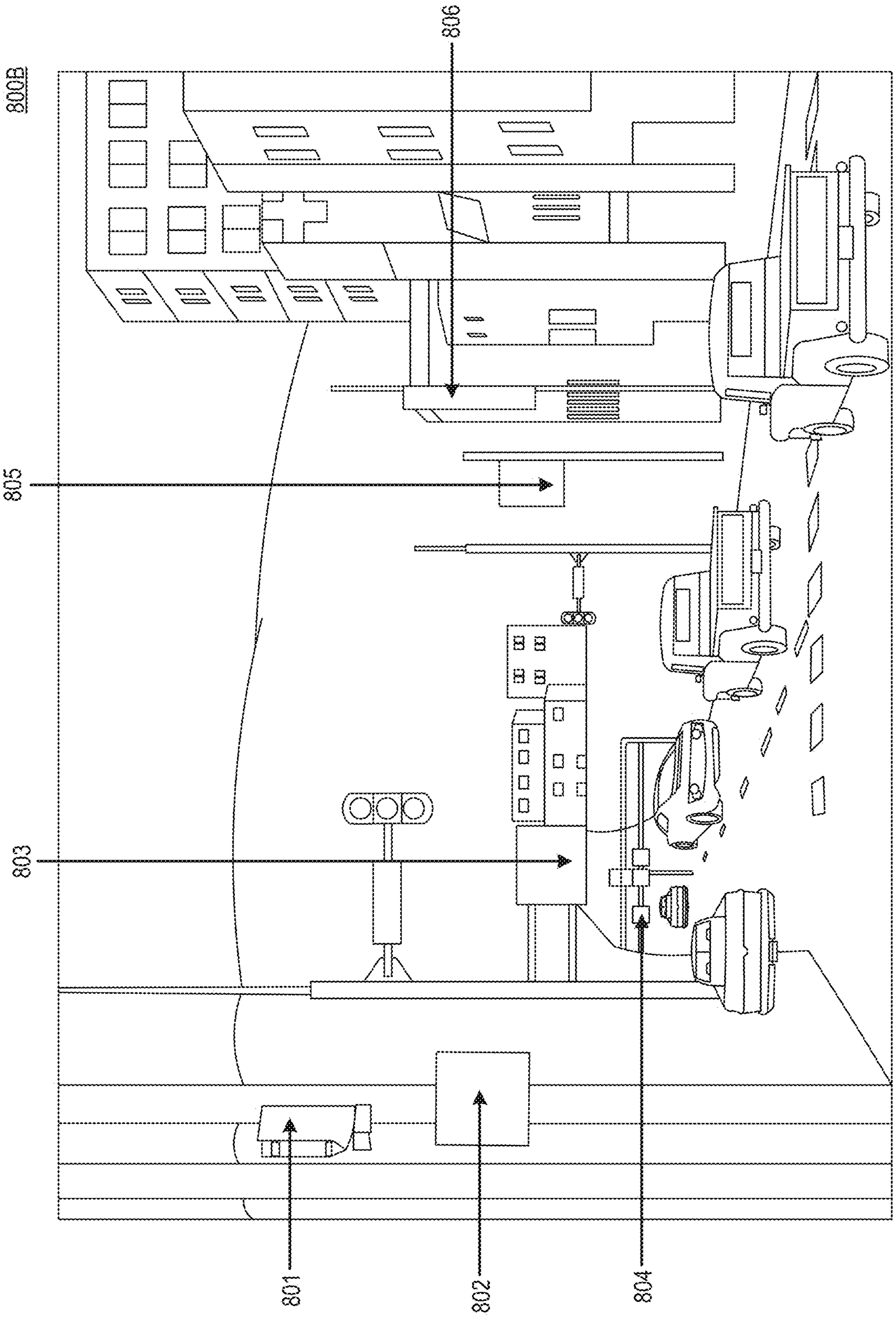


FIG. 8B

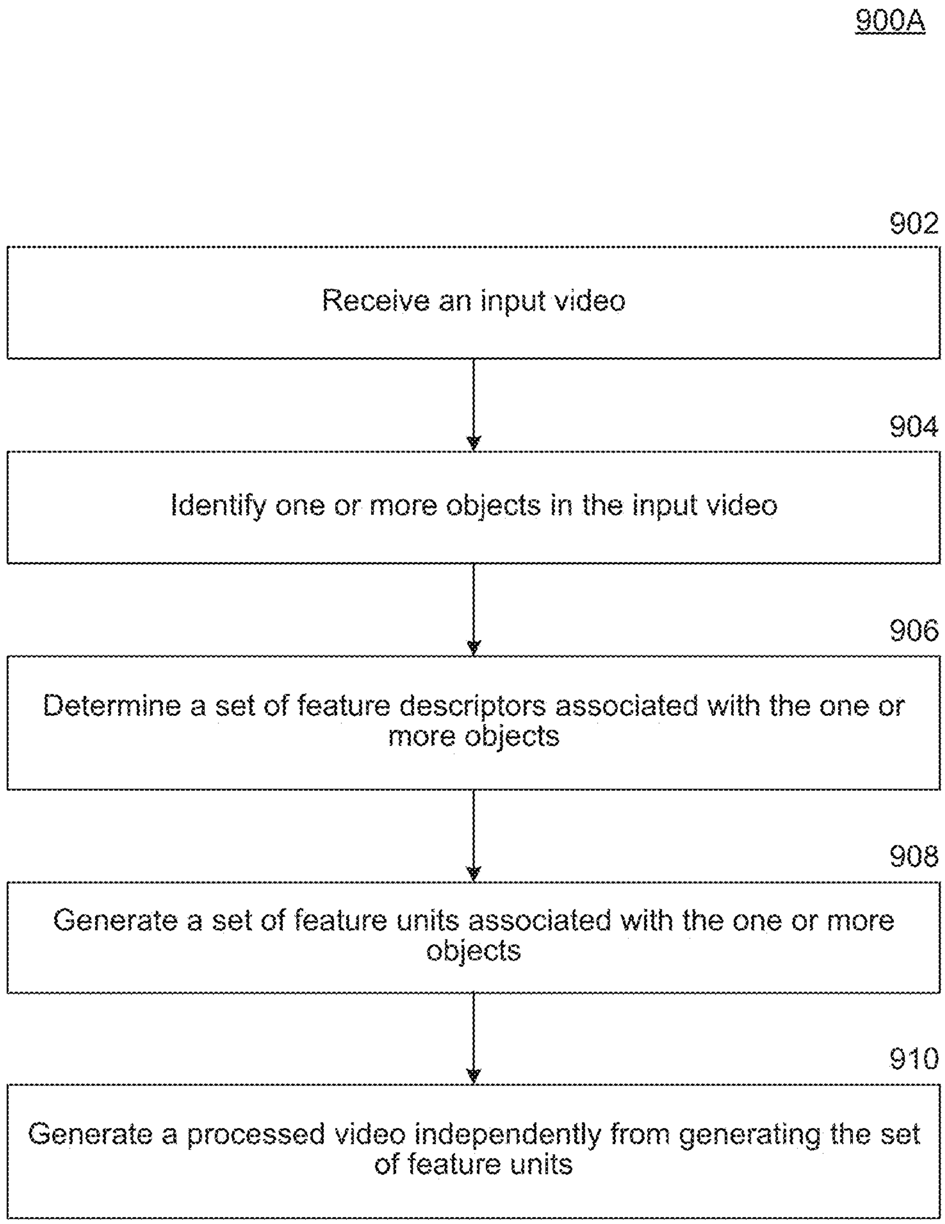

900A
902
Receive an input video
904
Identify one or more objects in the input video
906
Determine a set of feature descriptors associated with the one or more objects
908
Generate a set of feature units associated with the one or more objects
910
Generate a processed video independently from generating the set of feature units

FIG. 9A

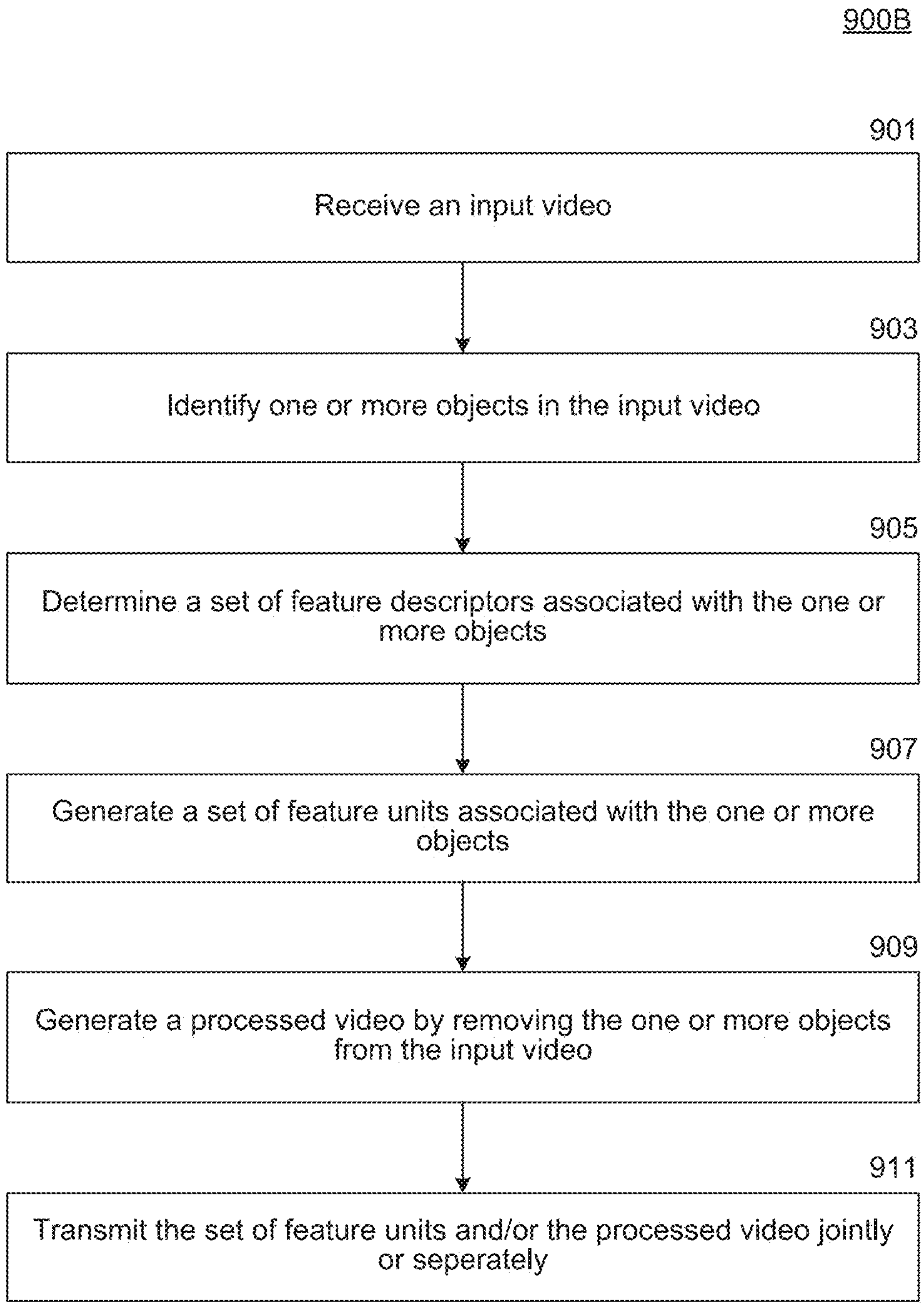

900B
901
Receive an input video
903
Identify one or more objects in the input video
905
Determine a set of feature descriptors associated with the one or more objects
907
Generate a set of feature units associated with the one or more objects
909
Generate a processed video by removing the one or more objects from the input video
911
Transmit the set of feature units and/or the processed video jointly or seperately

FIG. 9B

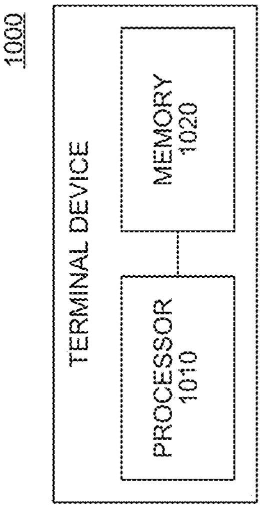


FIG. 10

METHOD AND SYSTEM FOR PROCESSING VIDEO

CROSS-REFERENCE TO RELATED APPLICATION(S)

[0001] This application is a continuation of International Application No. PCT/CN2022/077142, filed Feb. 21, 2022, which claims priority to European Patent Application No. 21461589.0, filed Sep. 13, 2021, the entire disclosures of which are hereby incorporated by reference.

TECHNICAL FIELD

[0002] This application relates to video coding and more particularly to a method and system for processing a video.

BACKGROUND

[0003] Video compression has been used for transmitting video data. Higher compression rates help ease the required resources for transmission but can result in loss of video qualities. For images observed by human beings, loss of image quality can affect aesthetic factors (e.g., "looks good" or not) of a video and accordingly deteriorate user experience. However, for images to be recognized by machines (e.g., self-driving vehicles), the context of the images is more important than the aesthetic factors of the video. Recent development of Video Coding for Machines (VCM) can be found in ISO/IEC JTC 1/SC 29/WG 2 N18 "Use cases and requirements for Video Coding for Machines." To reduce transmission time and consumption of transmission resources for video data for machines, it is desirable to have an improved system and method to effectively encode and decode the video data for machines.

SUMMARY

[0004] The present disclosure provides a method for processing a video. The method includes: receiving an input video; identifying one or more objects in the input video; determining a set of feature descriptors associated with the one or more objects; and generating a set of feature units associated with the one or more objects, where the set of feature units include Network Abstraction Layer (NAL) units.

[0005] The present disclosure further provides a system for processing a video. The system includes a transmitter configured to receive an input video; identify one or more objects in the input video; determine a set of feature descriptors associated with the one or more objects; generate a set of feature units associated with the one or more objects; and generate a processed video independently from generating the set of feature units.

BRIEF DESCRIPTION OF THE DRAWINGS

[0006] To describe the technical solutions in the implementations of the present disclosure more clearly, the following briefly describes the accompanying drawings. The accompanying drawings show merely some aspects or implementations of the present disclosure, and a person of ordinary skill in the art may still derive other drawings from these accompanying drawings without creative efforts.

[0007] FIG. **1** is a schematic diagram of a wireless communication system in accordance with one or more implementations of the present disclosure.

[0008] FIG. **2**A is a schematic diagram illustrating a system in accordance with one or more implementations of the present disclosure.

[0009] FIG. **2**B is a schematic diagram illustrating video coding based on feature descriptors and feature units in accordance with one or more implementations of the present disclosure. FIG. **2**B is only an example where the feature descriptors and the feature units can be transmitted in a bitstream. It should be understood that in other embodiments, the feature descriptors can be processed independently.

[0010] FIG. **3** is a schematic diagram illustrating a transmitter in accordance with one or more implementations of the present disclosure.

[0011] FIG. **4** is a schematic diagram illustrating a receiver in accordance with one or more implementations of the present disclosure.

[0012] FIG. **5** is a schematic diagram illustrating image processing of an object in accordance with one or more implementations of the present disclosure.

[0013] FIGS. **6**A, **6**B, **6**C and **6**D are examples of images processed by the methods in the present disclosure.

[0014] FIGS. **7**A and **7**B are examples of images processed by the methods in the present disclosure.

[0015] FIGS. **8**A and **8**B are examples of images processed by the methods in the present disclosure.

[0016] FIGS. **9**A and **9**B are a flowcharts illustrating methods in accordance with one or more implementations of the present disclosure.

[0017] FIG. **10** is a schematic block diagram of a terminal device in accordance with one or more implementations of the present disclosure.

DETAILED DESCRIPTION

[0018] The present disclosure provides apparatuses and methods for processing video data for machines. In some embodiments, the machines can include self-driving vehicles, robots, aircrafts, and/or other suitable devices or computing systems that are capable of video data processing and/or analysis, e.g. using artificial intelligence. More particularly, the present disclosure provides a method for video coding based on "feature descriptors" and "feature units" associated with one or more objects in a video. To reduce the size of transmitting the video, the video is processed to remove the objects from the video. To preserve the features of the removed objects, "feature descriptors" (e.g., decryption of the object such as: the object being a traffic/graphic sign or an advertisement sign, the object having a triangular shape, a size of the object, a boundary/location of the object,) and "feature units" (e.g., image units of the object, such as pixels, pictures, slices, tiles, etc. associated with the object) can be generated. These feature descriptors and feature units can be compressed/encoded and transmitted along with the processed video (with the objects removed). By this arrangement, the processed video can be encoded or compressed in a higher compression rate (which reduce the data size for transmission) without losing the features of the objects in the video. For example, the present disclosure provides high level syntax of a joint bitstream of "data" and "features." Descriptions of the "features" can be hierarchically arranged into synchronized structure of data units used in video coding. For example, in Advanced Video Coding (AVC), High Efficiency Video Coding (HEVC), and Versatile Video Coding (VVC), these data units can be in Network

Abstraction Layer (NAL). Although embodiments discussed herein can include video coding involving Compact Descriptors for Visual Search (CDVS), Compact Descriptors for Video Analysis (CDVA), classic scale-invariant feature transform (SIFT), etc., these embodiments are only non-limiting examples.

[0019] One aspect of the present disclosure is that it provides system and methods for transmitting a video based on feature descriptors and feature units associated with one or more objects in the video. In some embodiments, the feature descriptors and feature units can be customized (e.g., have a multiple level arrangement) by a system operator so as to effectively transmit the video. Embodiments discussing customizable feature descriptors and feature units are discussed in detail with reference to FIG. **2**B. The feature descriptors discussed in FIG. **2**B are only examples and are not limiting. In some embodiments, the feature descriptors can be independent from a bitstream (e.g., can be stored, analyzed, transmitted, etc. interpedently).

[0020] Another aspect of the present disclosure is that it enables an operator to set up multiple levels of feature descriptors. For example, the feature descriptors can include a higher-level descriptor that is indictive of a generic feature of an object in the video. Examples of the generic feature includes a type (e.g., traffic/graphic sign, advertisement sign, a logo, etc.) of the object, a shape of the object, etc. The feature descriptors can also include a lower-level descriptor that is indictive of a specific feature of the object. Examples of the specific feature includes a color, a size, a location, a boundary, a moving direction etc. of the object. In some embodiments, there can be only one level of feature descriptors. In some embodiments, there can be more than two levels of feature descriptors.

[0021] Another aspect of the present disclosure is that it provides a system for video coding for machines. The system can include (i) a transmitter configured to encode or compress video data based on identified objects and/or features of the video, and (ii) a receiver configured to decode or decompress the video data encoded or compressed by the foregoing transmitter.

[0022] When encoding a video, the transmitter can (i) identify one or more objects (e.g., a traffic sign, a road indicator, logos, tables, other suitable areas/fields that provide textual and/or numerical information, etc.) in the video; (ii) extract features (e.g., texts, numbers, and their corresponding colors, fonts, sizes, locations, etc.) associated with the identified objects; (iii) monitor and/or track the identified objects to determine or predict their moving directions and/or trajectories; (iv) processing images corresponding to the identified objects in each frame of the video (e.g., use a representative color to fill the whole area that the identified object occupies, so as to significantly reduce the resolution of that area); (v) encode (or compress) the video with the processed images; and (vi) transmit the encoded (or compressed) video and the extracted features via a network. In some embodiments, the video and the extracted objects (described by their features) may be transmitted in one bitstream or in multiple bitstreams. Embodiments of transmitting in one bitstream is discussed herein with referent to FIG. **2**B. Embodiments of the transmitter are discussed in detail with reference to FIG. **3**. In some embodiments, the extracted objects and corresponding feature descriptors can be independent from a bitstream (e.g., can be stored, analyzed, transmitted, etc. interpedently).

[0023] The present disclosure also provides a receiver configured to decode the encoded video. In some embodiments, the receiver can (a) receive an encoded video via a network; (b) decode the encode video based on identified objects and their corresponding features; (c) generate a decoded video with the identified objects. Embodiments of the receiver are discussed in detail with reference to FIG. **4**.

[0024] One aspect of the present disclosure is to provide methods for processing a video with objects. The method includes, for example, (1) identifying one or more objects in the video; (2) extracting features associated with the identified objects; (3) determining locations, moving directions, and/or trajectories of the identified objects; (4) processing the images corresponding to the identified objects in each frame of the video; (5) generating descriptors corresponding to the extracted features; (6) compressing the generated descriptors; (7) encoding the video with the processed images; (8) transmitting the encoded video and the compressed descriptors (e.g., by multiplexed bitstreams). In some embodiments, the method can further include (9) receiving the encoded video and the compressed descriptors via a network; (10) decompressing the compressed descriptors; and (11) decoding the encode video based on the decompressed descriptors. Embodiments the method are discussed in detail with reference to FIG. **5**.

[0025] Another aspect of the present disclosure is to provide methods for processing a video with objects. The method includes, for example, (i) receiving an input video; (ii) identifying one or more objects in the input video; (iii) determining a set of feature descriptors associated with the one or more objects; (iv) generating a set of feature units associated with the one or more objects; (v) generating a processed video by removing the one or more objects from the input video; and (vi) transmitting the set of feature descriptors, the set of feature units, and the processed video in a bit stream in a time period. Since the feature descriptors can be used to effectively represent the objects, the images of the objects can be removed from the input video and therefore do not need to be transmitted. Accordingly, the present method can effectively reduce the transmission size. For example, in the foregoing time period, assuming that there is a first number of the set of feature descriptors are transmitted in the time period and there is a second number of video units of the processed video are transmitted in the time period, the first number would be smaller than the second number. In other words, the present invention can transmit the video by using less transmission resources.

[0026] In some embodiments, the present method can be implemented by a tangible, non-transitory, computer-readable medium having processor instructions stored thereon that, when executed by one or more processors, cause the one or more processors to perform one or more aspects/features of the method described herein.

Communications Environment

[0027] FIG. **1** illustrates a system **100** for implementing the methods of the present disclosure. As shown in FIG. **1**, the system **100** includes a network device **101**. Examples of the network device **101** include a base transceiver station (Base Transceiver Station, BTS), a NodeB (NodeB, NB), an evolved Node B (eNB or eNodeB), a Next Generation NodeB (gNB or gNode B), a Wireless Fidelity (Wi-Fi) access point (AP), etc. In some embodiments, the network device **101** can include a relay station, an access point, an in-vehicle device, a wearable device, and the like. The network device **101** can include wireless connection devices for communication networks such as: a Global System for Mobile Communications (GSM) network, a Code Division Multiple Access (CDMA) network, a Wideband CDMA (WCDMA) network, an LTE network, a cloud radio access network (Cloud Radio Access Network, CRAN), an Institute of Electrical and Electronics Engineers (IEEE) 802.11-based network (e.g., a Wi-Fi network), an Internet of Things (IoT) network, a device-to-device (D2D) network, a next-generation network (e.g., a 5G network), a future evolved public land mobile network (Public Land Mobile Network, PLMN), or the like. A 5G system or network may be referred to as a new radio (New Radio, NR) system or network. In some embodiments, the present methods can be implemented in a wired communication system or a combination of wired or wireless systems. In some embodiments, the present methods can be implemented by stationary and/or mobile transmitter and/or receivers.

[0028] As shown in FIG. **1**, the system **100** also includes a terminal device **103**. The terminal device **103** can be an end-user device configured to facilitate wireless communication. The terminal device **103** can be configured to wirelessly connect to the network device **101** (via, e.g., a wireless channel **105**) according to one or more corresponding communication protocols/standards. The terminal device **103** may be mobile or fixed. The terminal device **103** can be a user equipment (UE), an access terminal, a user unit, a user station, a mobile site, a mobile station, a remote station, a remote terminal, a mobile device, a user terminal, a terminal, a wireless communications device, a user agent, or a user apparatus. Examples of the terminal device **103** include a modem, a cellular phone, a smart phone, a cordless phone, a Session Initiation Protocol (SIP) phone, a wireless local loop (WLL) station, a personal digital assistant (PDA), a handheld device having a wireless communication function, a computing device or another processing device connected to a wireless modem, an in-vehicle device, a wearable device, an IoT device, a terminal device in a future 5G network, a terminal device in a future evolved PLMN, or the like.

[0029] For illustrative purposes, FIG. **1** illustrates only one network device **101** and one terminal device **103** in the wireless communications system **100**. However, it is understood that, in some instances, the wireless communications system **100** can include additional/other devices, such as additional instances of the network device **101** and/or the terminal device **103**, a network controller, a mobility management entity/devices, etc.

[0030] In some embodiments, the network device **101** can act as a transmitter described herein. Alternatively, in some embodiments, the network device **101** can act as a receiver described herein. Similarly, the terminal device **103** can act as a transmitter described herein. Alternatively, in some embodiments, the terminal device **103** can act as a receiver described herein.

[0031] FIG. **2**A is a schematic diagram illustrating a system **200** in accordance with one or more implementations of the present disclosure. The system **200** includes a transmitter **201** and a receiver **203**. The transmitter **201** is configured to transmit encoded video data to the receiver **203** via a network **205**. The transmitter **201** includes a processor **2011**, a memory **2013**, and an encoder **2015**. In some embodiments, the transmitter **201** can be implemented as a chip (e.g., a system on chip, SoC). The processor **2011** is configured to implement the functions of the transmitter **201** and the components therein. The memory **2013** is configured to store data, instructions, and/or information associated with the transmitter **201**. The encoder **2015** is configured to process and encode a video with one or more objects. In some embodiments, the encoder **2015** can (1) identify one or more objects in the video; (2) extract one or more features associated with the one or more objects; (3) process each frame of the video based on the one or more objects; (4) encode the processed video; and (5) transmit the encoded video and the extracted feature.

[0032] In some embodiments, the encoder **2015** can be used to process video data for machines, such as vehicles, aircrafts, ships, robots, other suitable devices or computing systems that are capable of video data processing and/or analysis, e.g. using artificial intelligence. The encoder **2015** can first identify one or more objects in the video. Embodiments of the object can include, for example, a traffic sign, a road indicator, other suitable areas/fields that provide textual and/or numerical information, etc. In some embodiments, the object can be defined by a system operator (e.g., a particular shape, in a specific color, with certain textual features, etc.).

[0033] Once the object is identified, the encoder **2015** can extract one or more feature from the identified object. Examples of the extracted features include texts, numbers, and their corresponding colors, fonts, sizes, locations, etc. associated with the identified objects. For example, a traffic sign in a video can be identified as an object and the information "speed limit: 100 km/h" in the traffic sign can be the extracted feature. By separating the information carried by the traffic sign, the video including the traffic sign can be compressed in a higher ratio (which corresponds to a smaller data size for transmission), without worrying that doing so may result in the information become not recognizable due to the compression.

[0034] The encoder **2015** can further process the video by removing the images associated with the object in each frame of the video. In some embodiments, these images associated with the object can be replaced by a signal color (e.g., the same or similar to a surrounding image; a representative color, etc.) or a background image (e.g., a default background of a traffic sign). In some embodiments, these images can be left blank (to be generated by a decoder afterwards). The processed video (i.e., with the objects removed, replaced, or edited) can then be encoded (e.g., as a bitstream) for transmission.

[0035] In some embodiments, the encoder **2015** can be configured to track or monitor the identified objects such that it can determine or predict the locations, moving directions, and/or trajectories of the objects in the incoming frame. For example, the encoder **2015** can set a few locations (e.g., pixels) surrounding the objects as "check points" to track or monitor the possible location changes of the objects. By this arrangement, the encoder **2015** can effectively identify and manage the objects, without losing tracking of them. In some embodiments, information regarding the boundary of an object can be tracked and/or updated on a frame-by-frame basis.

[0036] The encoded video and the extracted feature can then be transmitted via the network **205**. In some embodiments, the encoded video and the extracted feature can be transmitted in two bitstreams. In some embodiments, the encoded video and the extracted feature can be transmitted in the same bitstream.

[0037] As shown in FIG. **2**A, the receiver **203** receives the encoded video and then can "restore" the encoded video by encoding it and adding the extracted feature thereto. For example, in some embodiments, the objects can be modified and added according to corresponding descriptors (e.g., viewing direction, a size/shape of the object, etc.) The receiver **203** includes a processor **2031**, a memory **2033**, and a decoder **2035**. In some embodiments, the receiver **203** can be implemented as a chip (e.g., a system on chip, SoC). The processor **2031** is configured to implement the functions of the receiver **203** and the components therein. The memory **2033** is configured to store data, instructions, and/or information associated with the receiver **203**. The decoder **2035** is configured to decode the encoded video and restore the removed/replaced objects therein. In some embodiments, the decoder **2035** can perform functions in a "reverse" fashion as the encoder **2015** to restore the video. In some embodiments the decoder **2035** can further process the video for better image quality or generate video based on user preferences.

[0038] In some embodiments, the transmitter **201** and the receiver **203** can both include an object database for storing reference object information (e.g., types of the objects; sample objects for comparison, etc.) for identifying the one or more objects. In some embodiments, the information stored in the object database can be trained by a machine learning process so as to enhance the accuracy of identifying the objects.

[0039] In some embodiments, the extracted feature can be described in a descriptor. The descriptor is indicative of the textual (e.g., a table of texts; road names, etc.), numerical (e.g., numbers shown), locational (a relative location of the object; a moving direction, etc.), contextual (the object is adjacent to a building or a road), and/or graphical (e.g., color, size, shape, etc.) information of the extracted feature. The descriptor can be stored in the object data database.

[0040] In some embodiments, the encoding, decoding, compressing and decompressing processes described herein can include coding processes involving Advanced Video Coding (AVC), High Efficiency Video Coding (HEVC), Versatile Video Coding (VVC), Alliance for Open Media Video 1 (AV1) or any other suitable methods, protocols, or standards.

[0041] FIG. **2**B is a schematic diagram illustrating video coding based on feature descriptors and feature units in accordance with one or more implementations of the present disclosure. FIG. **2**B includes a time frame **20** illustrating embodiments of multi-layer video coding of the present disclosure. As shown in FIG. **2**B, a video can be encoded and transmitted in five levels, higher-level feature descriptor (HFD) level **21**, lower-level feature descriptor (LFD) level **22**, feature unit (FU) level **23**, video unit (VU) level **24**, and video parameter set (VPS) level **25**. As indicated in FIG. **2**B, levels **21**-**23** are "features" levels, and levels **24** and **25** are "video" levels. In some embodiments, there can be only one level of feature descriptors. In other embodiments, there can be more than two levels of feature descriptors.

[0042] In the illustrated embodiments, during a period of time (from T**1** to T**5**), each level can have different transmission "frequencies." For example, during T**1**-T**5**, there is only one HFD transmitted (i.e., HFD at time T**1**). In the same time period, there are two LFDs (i.e., LFD**1** at T**1**; LFD**2** at T**3**) and two VPSs (i.e., VPS **1** at time T**1**; VPS**2** at T**2**.**5**) transmitted, and there are more than five FUs (i.e., FU**1** at time T**1**; FU**2** at T**2**; FU**3** at time T**3**; FU**4** at T**4**; FU**5** at T**5**) and VUs (i.e., VU**1** at time T**1**; VU**2** at T**2**; VU**3** at time T**3**; VU**4** at T**4**; VU**5** at T**5**) transmitted. In some embodiments, the video units VU**1**-**5** can be transmitted independent from the feature units FU**1**-**5**.

[0043] In the illustrated embodiments, the number of the transmitted HFD (i.e., only one) and LFDs (i.e., two) are less than the number of FUs (i.e., more than five) and VUs (i.e., more than five) transmitted. The present system can effectively use HFD, LFDs, and FUs to sufficiently describe features of the objects in the video.

[0044] The feature descriptors FDs are used to describe the features of the objects identified in the video. The higher-level feature descriptor HFD is indictive of a generic feature, such as a type (e.g., traffic sign, advertisement sign, a logo, etc.) of the object, a shape of the object, etc. The lower-level feature descriptor LFD is indictive of a specific feature of the object, such as a color, a size, a location, a boundary, a moving direction etc. of the object. The feature units FUs are the content of the features indicated by the feature descriptor. Examples of the FUs include pixels, pictures, slices, tiles, etc. associated with the object.

[0045] In some embodiments, the feature units FUs can be of various types and have different contents. In some embodiments, the feature units FUs can also be organized in layers. The feature units FUs (e.g., a traffic sign showing speed limits) correspond to video units VUs (e.g., an image having the traffic sign and captured by a front camera of a vehicle). The feature units FUs can be compressed or uncompressed. In some embodiments, the feature units FUs can be encoded using an "inter-unit" prediction (e.g., using interpolation schemes to determine a feature unit based on neighboring feature units).

[0046] In some embodiments, each feature unit corresponds to at least one parameter set (e.g., which can include one or more feature descriptors). In some embodiments, the feature parameter sets can be hierarchically organized.

[0047] The video units VUs are the video sequence of the video stream. The VUs are used to carry pictures, slices, etc., such as video coding layer (VCL) network abstraction layer unit (NALUs) in Advanced video Coding (AVC) or High Efficiency Video Coding (HEVC). The video parameter sets VPSs include information describing the video sequence, such as supplemental enhancement information (SEI), etc.

[0048] Based on the coding structure described in FIG. **2**B, the present system can effectively use HFD, LFDs, and FUs to sufficiently describe features of the objects in the video. By this arrangement, the features of the objects can be efficiently transmitted (e.g., it requires less transmission resources).

[0049] FIG. **3** is a schematic diagram illustrating a transmitter **300** in accordance with one or more implementations of the present disclosure. The transmitter **300** includes an object database **310**, an object recognition component **311**, a video processing component **312**, a compressing component **313**, a video encoder **314**, and a bitstream multiplexer (or a transmitting component) **315**. The foregoing components can be controlled or managed by a processor of the transmitter **300**.

[0050] When an input video **31** comes into the transmitter **300**, it can be directed to the object recognition component

311 and the video processing component **312**. In some embodiments, the input video **31** can come first to the object recognition component **311** and then to the video processing component **312**.

[0051] The object recognition component **311** is configured to recognize one or more objects in the video. As shown in FIG. **3**, the object recognition component **311** is coupled to the object database **310**. The object database **310** stores reference object information (e.g., types of the objects; sample objects for comparison, etc.) for identifying the one or more objects in a video. In some embodiments, the information stored in the object database **310** can be trained by a machine learning process so as to enhance the accuracy of identifying the objects. The object recognition component **311** can send a query and receive a query response **36** from the object database **310**. The query response **36** can facilitate the object recognition component **311** to identify and/or determine one or more objects in the input video **31**.

[0052] Once the one or more objects have been identified, one or more features associated with the one or more objects can be extracted. Examples of the extracted features include texts, numbers, and their corresponding colors, fonts, sizes, locations, etc. associated with the identified objects. One or more descriptors **34** can be generated based on the extracted features. The descriptors **34** are indicative of the foregoing features of the one or more objects (e.g., what the features are and where they are located, etc.). The descriptors **34** are sent to the video processing component **312** and the compressing component **313** for further process.

[0053] After the compressing component **313** receives the descriptors **34**, the descriptors **34** are compressed so as to generate compressed descriptors **35**. In some embodiments, the compression rate of the foregoing compression can be determined based on the content of the descriptors **34**. The compressed descriptors **35** is then sent to the bitstream multiplexer **315** for further process.

[0054] After the video processing component **312** receives the descriptors **34**, the input video is processed by removing the identified objects therein (e.g., based on the information provided by the descriptors **34**). The video processing component **312** then generates a processed video **32** (with the identified objects removed). In some embodiments, the removed object can be replaced by a blank, a background color, a background image, or a suitable item with lower image resolution than the removed objects. Embodiments of the blank, the background color, and the background image are discussed in detail with reference to FIG. **5**. The processed video **32** is then sent to the video encoder **314** for further process.

[0055] The video encoder **314** then encodes the processed video **32** by using a video coding scheme such as AVC, HEVC, VVC, AV1, or any other suitable methods, protocols, or standards. The video encoder **314** then generates an encoded video **33**, which is sent to the bitstream multiplexer **315** for further process.

[0056] After receiving the encoded video **33** and the compressed descriptors **35**, the bitstream multiplexer **315** can generate a multiplexed bitstream **37** for transmission. In some embodiments, the multiplexed bitstream **37** can include two bitstreams (i.e., one is for the encoded video **33**; the other is for the compressed descriptors **35**). In some embodiments, the multiplexed bitstream **37** can be a single bitstream. In some embodiments, the transmitter **300** can be implemented without the multiplexed bitstream **37**.

[0057] FIG. **4** is a schematic diagram illustrating a receiver **400** in accordance with one or more implementations of the present disclosure. The receiver **400** includes a bitstream demultiplexer (or a receiving component) **415**, an object description decoder **413**, an object reconstruction component **411**, an object database **410**, a video decoder **414**, and a video merging component **412**.

[0058] The bitstream demultiplexer **415** receives and multiplexes a multiplexed compressed bitstream **40**. Accordingly, the bitstream demultiplexer **415** can generate a compressed descriptors **41** and an encoded video **42**. The encoded video **42** is sent to the video decoder **414**. The video decoder **414** then decodes the encoded video **42** and generates decoded video **44** (with objects removed). The decoded video **44** is sent to the video merging component **412** for further process.

[0059] The compressed descriptors **41** is sent to the object description decoder **413**. The object description decoder **413** can decode the compressed descriptors **41** and then generates descriptors **43**. The descriptors **43** are indicative of one or more extracted features corresponding to one or more objects. The descriptors **43** are sent to the object reconstruction component **411** for further process.

[0060] The object reconstruction component **411** is coupled to the object database **410**. The object database **410** stores reference object information (e.g., types of the objects; sample objects for comparison, etc.) for recognizing the one or more objects corresponding based on the descriptors **43**. In some embodiments, the information stored in the object database **410** can be trained by a machine learning process so as to enhance the accuracy of identifying the objects. The object reconstruction component **411** can send a query and receive a query response **45** from the object database **410**. The query response **45** can facilitate the object reconstruction component **411** to recognize the one or more objects indicated by the descriptors **43**. Accordingly, the object reconstruction component **411** can generate reconstructed objects **46**. The reconstructed objects **46** can be sent to the video merging component **412** for further process. In some embodiments, the reconstructed objects **46** can also be sent and used for reference or machine-vision/machine-learning studies.

[0061] After receiving the reconstructed objects **46** and the decoded video **44**, the video merging component **412** merges the reconstructed objects **46** and the decoded video **44** and generates a decoded video with objects **47**. The decoded video with objects **47** has a resolution suitable for human beings (as well as machines) to recognize the objects therein.

[0062] FIG. **5** is a schematic diagram illustrating image processing of an object **501** in accordance with one or more implementations of the present disclosure. As shown in FIG. **5**, the object **501** is in an image **500**, which includes multiple grids **50**. When the object **501** is identified or recognized, the image **500** can be processed to remove the object **501** by performing an "over-the-hole" process. For example, during the "over-the-hole" process, the grids occupied by the object **501** can be removed. These grids can be replaced by a blank **503**, a background image **505**, or values interpolated by adjacent grids **51**, **52**. By this arrangement, the object **501** can be removed from the image **500** and an image **507** with object removed can be generated for process.

[0063] FIGS. **6**A, **6**B and **6**C are examples of images processed by the methods in the present disclosure. FIG. **6**A shows an original image **600**A includes a traffic sign **601**

(i.e., an object) therein. FIG. **6**B shows a processed image **600**B with the traffic sign **601** removed and replaced with background colors. The processed image **600**B can be transmitted with a high image compression rate without concerns of losing information carried by the traffic sign **601**. In some embodiment, the traffic sign **601** can only be "partially removed" by reducing its resolution, as shown in an processed image **600**C of FIG. **6**C. After transmission, the processed image **600**B or **600**C can be restored or "inpainted" as a restored image **600**D shown in FIG. **6**D.

[0064] FIGS. **7**A and **7**B are examples of images processed by the methods in the present disclosure. In the embodiments illustrated in FIGS. **7**A and **7**B, an original image **700**A includes two objects, a traffic sign **701** and a lane indicator **702**. As shown in FIG. **7**B, both objects can be removed as shown in a processed image **700**B.

[0065] In some embodiments, there can be more than two objects in an image. In FIG. **8**A, an original image **800**A includes multiple potential objects. The present system and methods enable an operator to determine the criteria to identify objects and determine whether to process the images associated the objects. For example, as shown in FIG. **8**B, six objects **801**-**806** are identified and processed. However, in other embodiments, the operator can determine only to process a portion of the identified objects (e.g., only process objects **801**-**804**, but not objects **805**, **806**). The present system enables the operator to customize the object identification process (e.g., what type of object is to be identified) and determine whether to process certain type of objects.

[0066] FIG. **9**A is flowchart illustrating a method **900**A in accordance with one or more implementations of the present disclosure. The method **900**A can be implemented by a system (the system **200**), a transmitter (e.g., the transmitter **201** or **300**), and/or a receiver (e.g., the receiver **203** or **400**) to process a video.

[0067] At block **902**, the method **900**A starts by receiving an input video.

[0068] At block **904**, the method **900**A continues to identify one or more objects in the video. Embodiments of the object can include, for example, a traffic sign, a road indicator, other suitable areas/fields that provide textual and/or numerical information, etc. In some embodiments, the object can be defined by a system operator (e.g., a particular shape, in a specific color, with certain textual features, etc.).

[0069] Optionally, the method **900**A may extract features associated with the identified objects. Examples of the extracted features include texts, numbers, and their corresponding colors, fonts, sizes, locations, etc. associated with the identified objects. For example, a traffic sign in a video can be identified as an object and the information "distance to City X" in the traffic sign can be the extracted feature.

[0070] At block **906**, the method **900**A continues to determine a set of feature descriptors associated with the one or more objects (e.g., corresponding to the extracted features). In some embodiments, the descriptor can be indicative of the textual (e.g., a table of texts; road names, etc.), numerical (e.g., numbers shown), locational (a relative location of the object; a moving direction, etc.), contextual (the object is adjacent to a building or a road), and/or graphical (e.g., color, size, shape, etc.) information of the extracted feature. The descriptor can be stored in the object data database. In some embodiments, the feature descriptors can include a result of one or more processes that are used to identify the objects, such as segmentation, classification, edge detection, corner detection, face detection etc.

[0071] At block **908**, the method **900**A continues to generate a set of feature units associated with the one or more objects. In some embodiments, the set of feature units can include Network Abstraction Layer (NAL) units, such as Visual Component Library units of non-VCL units described in AVC, HEVC, and/or VVC. In some embodiments, the set of feature units can include SIFT/CDVS feature units. In some embodiments, the set of feature units can include suitable units that can describe the one or more objects.

[0072] At block **910**, the method **900**A includes generating a processed video "independently" from generating the set of feature units. In some embodiments, the processed video can be generated from the input video by removing the video relevant to the identified objects. The processed video can be further filtered, compressed, or encoded. The processed video can include multiple video units. In some embodiments, the processed video and the set of feature units can be transmitted jointly, e.g., in a joint bitstream or multiple bitstreams. In some embodiments, the processed video and the set of feature units can be transmitted separately (e.g., in separate bitstreams).

[0073] In some embodiments, the processed video can include encoded units that are generated based on the set of feature units. For example, the set of feature unit can be embedded in the processed video.

[0074] The processed video and the set of feature units can be stored separately. The set of feature units can include a first number of feature units, and the processed video include a second number of video units. The first number and the second number do not need to be the same. In some embodiments, the first number can be smaller than the second number. By this arrangement, the present method can effectively and efficiently process and transmit the input video. For example, the input video can be transmitted in a higher compression ratio without losing the details of the objects.

[0075] FIG. **9**B is flowchart illustrating a method **900**B in accordance with one or more implementations of the present disclosure. The method **900**B can be implemented by a system (the system **200**), a transmitter (e.g., the transmitter **201** or **300**), and/or a receiver (e.g., the receiver **203** or **400**) to process a video.

[0076] At block **901**, the method **900**B starts by receiving an input video.

[0077] At block **903**, the method **900**B continues to identify one or more objects in the video. Embodiments of the object can include, for example, a traffic sign, a road indicator, other suitable areas/fields that provide textual and/or numerical information, etc. In some embodiments, the object can be defined by a system operator (e.g., a particular shape, in a specific color, with certain textual features, etc.).

[0078] Optionally, the method **900**B may extract features associated with the identified objects. Examples of the extracted features include texts, numbers, and their corresponding colors, fonts, sizes, locations, etc. associated with the identified objects. For example, a traffic sign in a video can be identified as an object and the information "speed limit: **100** km/h" in the traffic sign can be the extracted feature.

[0079] At block **905**, the method **900**B continues to determine a set of feature descriptors associated with the one or more objects. For example, this can be done by processing the images corresponding to the identified objects in each frame of the video. In some embodiments, the images can be processed by removing the objects therein (see, e.g., FIG. **5**). In some embodiments, the feature descriptors can be generated corresponding to the extracted features. In some embodiments, the feature descriptor can be indicative of the textual (e.g., a table of texts; road names, etc.), numerical (e.g., numbers shown), locational (a relative location of the object; a moving direction, etc.), contextual (the object is adjacent to a building or a road), and/or graphical (e.g., color, size, shape, etc.) information of the extracted feature. The descriptor can be stored in the object data database.

[0080] At block **907**, the method **900**B continues to generate a set of feature units associated with the one or more objects. At block **909**, the method **900**B continues to generate a processed video by removing the one or more objects from the input video. At block **911**, the method **900**B continues to transmit the set of feature units and/or the processed video jointly or separately.

[0081] In some embodiments, the set of feature units and the processed video can be transmitted in a joint bitstream or multiple bitstreams. In some embodiments, the set of feature units and the processed video can be transmitted and/or stored separately. In some embodiments, the set of feature units and the processed video can be compressed in the same scheme or different schemes.

[0082] In some embodiments, the encoded video and the descriptors can be multiplexed and transmitted in a single bitstream or two bitstreams. In some embodiments, the method **900** can further include receiving the encoded video and the compressed descriptors via a network; decompressing the compressed descriptors; and decoding the encode video based on the decompressed descriptors.

[0083] FIG. **10** is a schematic block diagram of a terminal device **1000** (e.g., an example of the terminal device **103** of FIG. **1**) in accordance with one or more implementations of the present disclosure. As shown in FIG. **10**, the terminal device **1000** includes a processing unit **1010** and a memory **1020**. The processing unit **1010** can be configured to implement instructions that correspond to the terminal device **1000**.

[0084] It should be understood that the processor in the implementations of this technology may be an integrated circuit chip and has a signal processing capability. During implementation, the steps in the foregoing method may be implemented by using an integrated logic circuit of hardware in the processor or an instruction in the form of software. The processor may be a general-purpose processor, a digital signal processor (DSP), an application specific integrated circuit (ASIC), a field programmable gate array (FPGA) or another programmable logic device, a discrete gate or transistor logic device, and a discrete hardware component. The methods, steps, and logic block diagrams disclosed in the implementations of this technology may be implemented or performed. The general-purpose processor may be a microprocessor, or the processor may be alternatively any conventional processor or the like. The steps in the methods disclosed with reference to the implementations of this technology may be directly performed or completed by a decoding processor implemented as hardware or performed or completed by using a combination of hardware and software modules in a decoding processor. The software module may be located at a random-access memory, a flash memory, a read-only memory, a programmable read-only memory or an electrically erasable programmable memory, a register, or another mature storage medium in this field. The storage medium is located at a memory, and the processor reads information in the memory and completes the steps in the foregoing methods in combination with the hardware thereof.

[0085] It may be understood that the memory in the implementations of this technology may be a volatile memory or a non-volatile memory, or may include both a volatile memory and a non-volatile memory. The non-volatile memory may be a read-only memory (ROM), a programmable read-only memory (PROM), an erasable programmable read-only memory (EPROM), an electrically erasable programmable read-only memory (EEPROM) or a flash memory. The volatile memory may be a random-access memory (RAM) and is used as an external cache. For exemplary rather than limitative description, many forms of RAMs can be used, and are, for example, a static random-access memory (SRAM), a dynamic random-access memory (DRAM), a synchronous dynamic random-access memory (SDRAM), a double data rate synchronous dynamic random-access memory (DDR SDRAM), an enhanced synchronous dynamic random-access memory (ESDRAM), a synchronous link dynamic random-access memory (SLDRAM), and a direct Rambus random-access memory (DR RAM). It should be noted that the memories in the systems and methods described herein are intended to include, but are not limited to, these memories and memories of any other suitable type.

[0086] The above Detailed Description of examples of the disclosed technology is not intended to be exhaustive or to limit the disclosed technology to the precise form disclosed above. While specific examples for the disclosed technology are described above for illustrative purposes, various equivalent modifications are possible within the scope of the described technology, as those skilled in the relevant art will recognize. For example, while processes or blocks are presented in a given order, alternative implementations may perform routines having steps, or employ systems having blocks, in a different order, and some processes or blocks may be deleted, moved, added, subdivided, combined, and/or modified to provide alternative implementations or subcombinations. Each of these processes or blocks may be implemented in a variety of different ways. Also, while processes or blocks are at times shown as being performed in series, these processes or blocks may instead be performed or implemented in parallel, or may be performed at different times. Further, any specific numbers noted herein are only examples; alternative implementations may employ differing values or ranges.

[0087] In the Detailed Description, numerous specific details are set forth to provide a thorough understanding of the presently described technology. In other implementations, the techniques introduced here can be practiced without these specific details. In other instances, well-known features, such as specific functions or routines, are not described in detail in order to avoid unnecessarily obscuring the present disclosure. References in this description to "an implementation/embodiment," "one implementation/embodiment," or the like mean that a particular feature, structure, material, or characteristic being described is included in at least one implementation of the described technology. Thus, the appearances of such phrases in this specification do not necessarily all refer to the same implementation/embodiment. On the other hand, such references are not necessarily mutually exclusive either. Furthermore, the particular features, structures, materials, or characteristics can be combined in any suitable manner in one or more implementations/embodiments. It is to be understood that the various implementations shown in the figures are merely illustrative representations and are not necessarily drawn to scale.

[0088] Several details describing structures or processes that are well-known and often associated with communications systems and subsystems, but that can unnecessarily obscure some significant aspects of the disclosed techniques, are not set forth herein for purposes of clarity. Moreover, although the following disclosure sets forth several implementations of different aspects of the present disclosure, several other implementations can have different configurations or different components than those described in this section. Accordingly, the disclosed techniques can have other implementations with additional elements or without several of the elements described below.

[0089] Many implementations or aspects of the technology described herein can take the form of computer- or processor-executable instructions, including routines executed by a programmable computer or processor. Those skilled in the relevant art will appreciate that the described techniques can be practiced on computer or processor systems other than those shown and described below. The techniques described herein can be implemented in a special-purpose computer or data processor that is specifically programmed, configured, or constructed to execute one or more of the computer-executable instructions described below. Accordingly, the term "processor" as generally used herein refers to any data processor. Information handled by the processors can be presented at any suitable display medium. Instructions for executing computer- or processor-executable tasks can be stored in or on any suitable computer-readable medium, including hardware, firmware, or a combination of hardware and firmware. Instructions can be contained in any suitable memory device, including, for example, a flash drive and/or other suitable medium.

[0090] The terms "coupled" and "connected," along with their derivatives, can be used herein to describe structural relationships between components. It should be understood that these terms are not intended as synonyms for each other. Rather, in particular implementations, "connected" can be used to indicate that two or more elements are in direct contact with each other. Unless otherwise made apparent in the context, the term "coupled" can be used to indicate that two or more elements are in either direct or indirect (with other intervening elements between them) contact with each other, or that the two or more elements cooperate or interact with each other (e.g., as in a cause-and-effect relationship, such as for signal transmission/reception or for function calls), or both. The term "and/or" in this specification is only an association relationship for describing the associated objects, and indicates that three relationships may exist, for example, A and/or B may indicate the following three cases: A exists separately, both A and B exist, and B exists separately.

[0091] These and other changes can be made to the disclosed technology in light of the above Detailed Description. While the Detailed Description describes certain examples of the disclosed technology, as well as the best mode contemplated, the disclosed technology can be practiced in many ways, no matter how detailed the above description appears in text. Details of the system may vary considerably in its specific implementation, while still being encompassed by the technology disclosed herein. As noted above, particular terminology used when describing certain features or aspects of the disclosed technology should not be taken to imply that the terminology is being redefined herein to be restricted to any specific characteristics, features, or aspects of the disclosed technology with which that terminology is associated. Accordingly, the invention is not limited, except as by the appended claims. In general, the terms used in the following claims should not be construed to limit the disclosed technology to the specific examples disclosed in the specification, unless the above Detailed Description section explicitly defines such terms.

[0092] A person of ordinary skill in the art may be aware that, in combination with the examples described in the implementations disclosed in this specification, units and algorithm steps may be implemented by electronic hardware, or a combination of computer software and electronic hardware. Whether the functions are performed by hardware or software depends on particular applications and design constraint conditions of the technical solutions. A person skilled in the art may use different methods to implement the described functions for each particular application, but it should not be considered that the implementation goes beyond the scope of this application.

1. A method for processing a video, comprising:
receiving an input video;
identifying one or more objects in the input video;
determining a set of feature descriptors associated with the one or more objects; and
generating a set of feature units associated with the one or more objects, wherein the set of feature units include Network Abstraction Layer (NAL) units.

2. The method of claim **1**, further comprising:
generating a processed video independently from generating the set of feature units.

3. The method of claim **2**, further comprising:
transmitting the processed video and the set of feature units in a joint bitstream.

4. The method of claim **2**, further comprising:
transmitting the processed video and the set of feature units separately.

5. The method of claim **2**, further comprising:
storing the processed video and the set of feature units separately.

6. The method of claim **2**, wherein the processed video includes encoded units generated based on the set of feature units.

7. The method of claim **1**, wherein the NAL units include Visual Component Library (VCL) units.

8. The method of claim **1**, wherein the NAL units include non-VCL units.

9. The method of claim **8**, wherein:
the set of feature units include a first number of feature units;
the processed video include a second number of video units; and
the first number is different from the second number.

10. The method of claim **9**, wherein the first number is smaller than the second number.

11. A system for processing a video, comprising:

a transmitter configured to:

receive an input video;

identify one or more objects in the input video;

determine a set of feature descriptors associated with the one or more objects;

generate a set of feature units associated with the one or more objects; and

generate a processed video independently from generating the set of feature units.

12. The system of claim **11**, wherein the transmitter is further configured to:

transmit the processed video and the set of feature units in a joint bitstream.

13. The system of claim **11**, wherein the transmitter is further configured to:

transmit the processed video and the set of feature units separately.

14. The system of claim **11**, wherein the set of feature units include Network Abstraction Layer (NAL) units.

15. The system of claim **11**, wherein:

the set of feature units include a first number of feature units;

the processed video include a second number of video units; and

the first number is smaller than the second number.

* * * * *