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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
03 October 2024 (03.10.2024)

WIPO | PCT

(10) International Publication Number
WO 2024/201124 A1

(51) International Patent Classification:
H04N 21/234 (2011.01) *G06V 20/40* (2022.01)
G06V 10/70 (2022.01)

(21) International Application Number:
PCT/IB2023/056504

(22) International Filing Date:
23 June 2023 (23.06.2023)

(25) Filing Language: English

(26) Publication Language: English

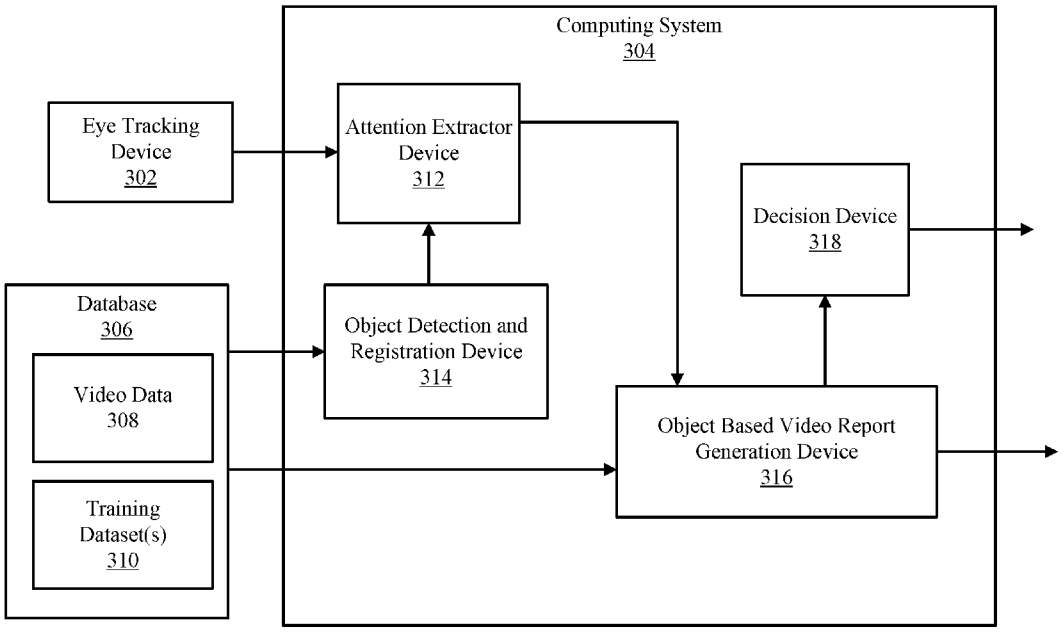
(30) Priority Data:
63/455,611 30 March 2023 (30.03.2023) US

(71) Applicant: **NEC LABORATORIES EUROPE GMBH** [DE/DE]; Kurfuersten-Anlage 36, 69115 Heidelberg (DE).

(72) Inventors: **SHALINI, Shalini**; c/o NEC Laboratories Europe GmbH, Kurfuersten-Anlage 36, 69115 Heidelberg (DE). **GONG, Na**; c/o NEC Laboratories Europe GmbH, Kurfuersten-Anlage 36, 69115 Heidelberg (DE).

(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, MG, MK, MN, MU, 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, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(54) **Title:** EYE-TRACKING-BASED OBJECT-FOCUSED VIDEO DESCRIPTION SYSTEM



300

FIG. 3

(57) **Abstract:** A method for generating attention based video description using eye-tracking includes obtaining, using an eye-tracker device, raw gaze data associated with a user watching at least a portion of video data comprising a plurality of frames. The method further includes identifying attention objects and extracting one or more frames from the plurality of frames. The method also includes generating one or more individual textual reports for each of the identified attention objects based on the one or more extracted frames and outputting the one or more second individual textual reports that describe the video data in context of each of the identified attention objects. In some embodiments, the one or more individual textual reports are generated based on optimizing generated descriptions for the one or more individual textual reports by utilizing a name entity recognition model and a deep learning-based image semantic segmentation model.

WO 2024/201124 A1

[Continued on next page]

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

EYE-TRACKING-BASED OBJECT-FOCUSED VIDEO DESCRIPTION SYSTEM

CROSS-REFERENCE TO PRIOR APPLICATION

[0001] Priority is claimed to U.S. Provisional Application No. 63/455,611, filed on March 30, 2023, the entire contents of which is hereby incorporated by reference herein.

FIELD

[0002] The present invention relates to a method, system and computer-readable medium for eye-tracking-based object-focused description of videos.

BACKGROUND

[0003] People spend a lot of time manually interpreting and describing the ample quantities of videos (e.g., closed-circuit television (CCTV) videos) that are available. Although it may not be possible to even realize how many parallel processes are active in people's brains when watching a video, since their brains are working so hard, people can recognize, recollect, think about, remember, and feel different emotions. Although insights and other advantages can be obtained by being able to figure out how to rapidly summarize relevant footage from the viewer's perspective, this is technically challenging and existing attention-based video summarization technology using eye-tracking approaches have limitations, especially for extremely long videos as users are compelled to watch the entire video.

SUMMARY

[0004] In an embodiment, the present disclosure provides a method for generating attention based video description using eye-tracking. Raw gaze data associated with a user watching at least a portion of video data comprising a plurality of frames is obtained using an eye-tracker device. Based on the raw gaze data, attention objects associated with objects of interest within the video data is identified. Further, one or more frames from the plurality of frames is extracted based on the identified attention objects. Then, one or more individual textual reports for each of the identified attention objects is generated based on the one or more extracted frames. Subsequently, the one or more individual textual reports are outputted. The one or more individual textual reports describe the video data in context of each of the identified attention objects. In some embodiments, the one or more individual textual reports are generated based on optimizing generated descriptions for the one or more individual textual reports by utilizing a name entity recognition model and a deep learning-based image semantic segmentation model.

BRIEF DESCRIPTION OF THE DRAWINGS

[0005] Embodiments of the present invention will be described in even greater detail below based on the exemplary figures. The present invention is not limited to the exemplary embodiments. All features described and/or illustrated herein can be used alone or combined in

different combinations in embodiments of the present invention. The features and advantages of various embodiments of the present invention will become apparent by reading the following detailed description with reference to the attached drawings which illustrate the following:

[0006] FIG. 1 schematically illustrates a method and system architecture for eye-tracking-based object-focused video description according to an embodiment of the present invention;

[0007] FIG. 2 illustrates an example dictionary for object attention extraction as an intermediate step according to an embodiment of the present invention;

[0008] FIG. 3 schematically illustrates an environment for eye-tracking-based object-focused video description according to an embodiment of the present invention; and

[0009] FIG. 4 is a block diagram of an exemplary processing system, which can be configured to perform any and all operations disclosed herein.

DETAILED DESCRIPTION

[0010] Embodiments of the present invention overcome limitations of existing video description technology by utilizing eye-tracking data to extract user-focused objects, and generating a more consistent and reliable description from a focused object's perspective using a self-optimized video descriptor.

[0011] Although existing video summarization approaches can provide concise summaries of videos that convey a key idea from the original, it has been recognized that existing technology suffers at least from the following limitations:

1. Ignore the psychological aspect of cognitive automation (e.g., what the watcher perceives from the video, and/or where he/she puts his/her attention).
2. Create a general summary based on the context of the full video rather than focusing on the objects in the video that the user is really interested in, and cannot describe the video from a focused object's perspective.
3. Methods based on key frame extraction that sum up the entire video using a few key frames or video skims that can result in information missing in the output.
4. Require watching the video completely to extract the key frames based on viewer's perspective.

[0012] Embodiments of the present invention overcome the aforementioned limitations by utilizing eye-tracking data of the viewer (e.g., the user) to identify attention objects from the video and create descriptions of the video based on the objects the viewer is focusing on, even if the viewer does not watch the video completely. Some existing eye-tracking based approaches generate summaries using key frames that have been retrieved from the video (see e.g., Bezugam, Sai Sukruth, et al., "Efficient Video Summarization Framework using EEG and Eye-tracking Signals," arXiv:2101.11249v1 (2021), which is hereby incorporated by reference

herein). In contrast, embodiments of the present invention enhance the computer functionality to enable generating the description based on the object on which viewer is giving attention. Embodiments of the present invention can catch attention in the beginning of what a viewer is looking at and describe the video completely from the object's perspective. For instance, embodiments of the present invention can first identify the attention object where the viewer places their attention on and then describe the video completely based on the attention object's perspective. Embodiments of the present invention can extract the attention object in the beginning of the video and describe the video completely. Thus, it might not be required to watch the full video to generate the report.

[0013] Embodiments of the present invention can also be practically applied to effect further improvements in a variety of technical fields, in addition to video analysis, including automated public safety and law enforcement tools (e.g., for policing), performance analysis (e.g., in sports), and automated healthcare and medicine. For instance, in policing, the CCTV can be thoroughly summarized in a so-called Manual of Guidance 5 (MG5) report in every charge and summons case if it serves as essential evidence. When a police officer pays attention to a suspect when watching the video, an embodiment of the present invention can automatically fill out an MG5 report and produce a description of what that individual does in the video. Additionally, and/or alternatively, if police wants to track a suspect or criminal, embodiments of the present invention enable the easy generation of a report for a specific person in a crowd. As a further advantage, embodiments of the present invention can be used to reduce the psychological burden on police officers, e.g., while having to watch sex assaults or domestic violence videos, by generating the description of the video based on a police officer's attention object even if the police officer only watches it for a short time (e.g., a few minutes). Embodiments of the present invention can catch attention in the beginning of what the police officer is looking at and overcome human effort by describing it fully, while avoiding the psychological burden of having to watch the entire video.

[0014] According to a first aspect, the present disclosure provides a method for generating attention based video description using eye-tracking. The method includes obtaining, using an eye-tracker device, raw gaze data associated with a user watching at least a portion of video data comprising a plurality of frames. The method further includes identifying, based on the raw gaze data, attention objects associated with objects of interest within the video data and extracting one or more frames from the plurality of frames based on the identified attention objects. The method also includes generating one or more individual textual reports for each of the identified attention objects based on the one or more extracted frames and outputting the one or more

individual textual reports that describe the video data in context of each of the identified attention objects.

[0015] According to a second aspect, the method according to the first aspect further comprises receiving user input from one or more input devices indicating user indicated interested objects from within the video data, and identifying the attention objects is based on the user input.

[0016] According to a third aspect, the method according to the first or the second aspect further comprises pre-processing the raw gaze data to extract one or more eye tracking features associated with the user watching the video data and using one or more object detection models and the video data to generate annotated video frames indicating video objects and unique identifiers associated with the video objects. The one or more eye tracking features comprise one or more fixations that indicate one or more durations of when the user's sight is fixed on one or more stationary objects from the video data. Further, identifying the attention objects is based on the one or more extracted eye tracking features and the annotated video frames.

[0017] According to a fourth aspect, the method according to any of the first to third aspects further comprises that determining one or more tags for each of the video objects based on fixation overlap between the one or more extracted eye tracking features and the annotated video frames, generating a dictionary based on the one or more tags for the video objects, comparing the one or more tags for each of the video objects with a fixation threshold and an attention threshold, and classifying one or more video objects of the video objects as the attention objects based on the comparison. Additionally, the one or more tags indicate whether the user paid attention to the video objects while watching the video data.

[0018] According to a fifth aspect, the method according to any of the first to fourth aspects further comprises that generating the one or more individual textual reports for each of the identified attention objects is further based on optimizing generated descriptions for the one or more individual textual reports by utilizing a name entity recognition model and a deep learning-based image semantic segmentation model.

[0019] According to a sixth aspect, the method according to any of the first to fifth aspects further comprises that generating one or more individual textual reports for each of the identified attention objects comprises: generating one or more first individual textual reports for each of the identified attention objects based on the one or more extracted frames; and generating one or more second individual textual reports for each of the identified attention objects based on the first individual textual reports and a consistency degree. The one or more second individual textual reports indicate one or more object identifiers associated with the identified attention objects and video descriptions of the identified attention object throughout the video data, and

outputting the one or more individual textual reports comprises outputting the one or more second individual textual reports..

[0020] According to a seventh aspect, the method according to any of the first to sixth aspects further comprises that generating the one or more second individual textual reports for each of the identified attention objects comprises: inputting the one or more extracted frames associated with the identified attention objects into the deep learning-based image semantic segmentation model to generate labels for the one or more extracted frames; identifying named entities from the one or more first individual textual reports using the name entity recognition model; generating the consistency degree based on the named entities, the generated labels, and an entity consistency optimizer neural network; and generating the one or more second individual textual reports based on optimizing a video description model using the consistency degree and a loss function associated with the video description model.

[0021] According to an eighth aspect, the method according to any of the first to seventh aspects further comprises that the one or more second individual textual reports comprises a plurality of second individual textual reports for a plurality of attention objects, each of the plurality of second individual textual reports is associated with a single attention object, of the plurality of attention objects, and a video description for the single attention object. Further, outputting the one or more second individual textual reports comprises: ranking the plurality of attention objects; and providing for display a top ranked attention object from the plurality of attention objects.

[0022] According to an ninth aspect, the method according to any of the first to eighth aspects further comprises that obtaining the raw gaze data comprises generating, using an eye-tracking module for the eye-tracker device, the raw gaze data based on the user watching the video data, and pre-processing the raw gaze data comprises extracting, using a gaze data preprocessor module, the one or more eye tracking features. The one or more eye tracking features comprise the one or more fixations and one or more saccades. The one or more saccades indicate rapid eye movement of the user when shifting focus from a first stationary object to a second stationary object.

[0023] According to a tenth aspect, the method according to any of the first to ninth aspects further comprises that the one or more object detection models is stored in an object detection and registration module, the one or more object detection models is an object detection deep neural network that is trained based on using training data or a pre-trained model that identifies the video objects, and the one or more object detection models identifies the video objects in each video frame of the plurality of frames and registers the video objects for unique identification.

[0024] According to an eleventh aspect, the method according to any of the first to tenth aspects further comprises that identifying the attention objects is based on using an object attention attractor module, and determining the attention objects based on the dictionary comprises determining the attention objects within their corresponding frames of the video data based on the dictionary and a fixation threshold.

[0025] According to a twelfth aspect, the method according to any of the first to eleventh aspects further comprises training the deep learning-based image semantic segmentation model to identify labels within frames of a video using one or more segmentation training datasets, and training a neural network based entity extractor to generate the name entity recognition model. The one or more segmentation training datasets comprises a MICROSOFT Common Objects in Context (MS-COCO) dataset. The neural network based entity extractor determines patterns of entities from given documents to identify important entities within the given documents, and generating the one or more second individual textual reports is based on using the deep learning-based image segmentation model and the name entity recognizer.

[0026] According to a thirteenth aspect, the method according to any of the first to twelfth aspects further comprising training a video summarization model using a domain specific dataset to generate a video description module and learning a similarity function between two entity sets using an entity consistency optimizer. The video summarization model describes sequences of frames corresponding to objects within the sequence of frames, and the entity consistency optimizer is a Multi-Player Perception (MLP) neural network.

[0027] A fourteenth aspect of the present disclosure provides a system for generating attention based video description using eye-tracking, the system comprising one or more hardware processors, which, alone or in combination, are configured to provide for execution of the following steps: obtaining, using an eye-tracker device, raw gaze data associated with a user watching video data comprising a plurality of frames; identifying, based on the raw gaze data, attention objects associated with objects of interest within the video data and extracting one or more frames from the plurality of frames based on the identified attention objects; generating one or more individual textual reports for each of the identified attention objects based on the one or more extracted frames; and outputting the one or more individual textual reports that describe the video data in context of each of the identified attention objects.

[0028] A fifteenth aspect of the present disclosure provides a tangible, non-transitory computer-readable medium having instructions thereon, which, upon being executed by one or more processors, provides for execution of the method according to any of the first to the thirteenth aspects.

[0029] FIG. 1 schematically illustrates a method and system architecture for eye-tracking-based object-focused video description according to an embodiment of the present invention. For instance, the method and system architecture 100 illustrates a comprehensive procedure and system architecture for how the attention-based video description using eye-tracking system automatically describes videos based on a user's perspective. In FIG. 1, inputs and modules (e.g., processors, engines, software instructions, controllers, hardware devices, and/or other computing apparatuses) are numbered 102-138. For instance, a sport trainer may use the video description system according to an embodiment of the present invention to view a video of a match and concentrate on one player to describe how that player performed. In the following, the system architecture 100 is described systematically in reference to the input and module numbers.

[0030] Step 1. Trigger the devices / system (e.g., the user 102, the video data 104, the eye-tracking module 106): For example, a user 102 starts the process by interacting with the eye-tracking module 106, and by providing a video of interest 104. The video can also be provided by another user or system.

[0031] For instance, the eye-tracking module 106 (e.g., eye-tracking device and/or eye-tracking equipment) can include a display device that displays a video using the video data 104. The video data 104 includes frames such as a first frame ($frame_1$), a second frame (f_2), and so on until a second to last frame (f_{n-1}) and a last frame (f_n). Additionally, and/or alternatively, a display device separate from the eye-tracking module 106 displays the video data 104. Another user and/or system (e.g., a system separate from the system architecture 100) can provide the video data 104. At step 1, the video data 104 is displayed on the display device frame-by-frame and the eye-tracking module 106 tracks the user 102 (e.g., the eyes of the user).

[0032] Step 2. Eye-tracking module 106: The system architecture 100 is designed to (e.g., configured to) be compatible with eye-trackers that provide gaze coordinates in real time, such as eye-tracker bars, glasses etc. The system architecture 100 takes this gaze data as input, which can be collected using any eye-tracker equipment, e.g., TOBII PRO-GLASSES-2 with variable sampling frequency depending on the eye-tracker (e.g., 60 hertz (Hz) for 100 milliseconds (ms)). Constraints related to display size, resolution, and distance of the subject to the screen are dependent solely on the technical restrictions of the hardware employed.

[0033] For instance, during the display of the video data 104, the eye-tracking module 106 tracks the eyes of the user 102 such as the location (e.g., gaze data) of where the user is watching. For example, the eye-tracking module 106 detects the location (e.g., the coordinates or the gaze coordinates) on the display device where the user 102 is watching, and provides the raw gaze data indicating the gaze coordinates to the attention extractor 112. For instance, for the

frames of the video data 104, the eye-tracking module 106 detects the gaze coordinates where the user 102 is watching, generates gaze data based on the gaze coordinates, and provides information 140 (e.g., the raw gaze data) to the attention extractor 112.

[0034] Step 3. Gaze data pre-processor 114: After collecting the user's eye-tracking data (e.g., the raw gaze data) through the eye-tracking module 106 (e.g., the eye-tracking equipment), the system architecture 100 applies the gaze data pre-processor 114 to pre-process this gaze data. For example, the eye-tracking module 106 provides raw gaze data 140 from which eye-tracking features such as fixations and/or saccades are extracted. For instance, the eye-tracking module 106 provides raw gaze data, which can be used by the gaze data preprocessor 114 to extract eye tracking features such as saccades and fixations. These features can further be used by the object based video report generation module 120. Fixation is the duration of an interval when a person's (e.g., user's 102) sight is fixed on a stationary object. A saccade is a rapid eye movement that occurs when a person (e.g., user 102) shifts their focus to another item. If the sampling rate of the eye-tracker is, e.g., 60 Hz, that means each timestamp (e.g., 1 second (sec)) corresponds to 60 points in the raw data obtained from the eye-tracking equipment. It is possible to down sample the data to map to the video frames if required. Apart from the gazing information collected via eye-tracking module 106 (e.g., the eye-tracking equipment), users (e.g., users 102) can provide their attention via various alternative inputs. For instance, users eye-tracking module 106 (e.g., the eye-tracking equipment) can tell the system architecture 100 their interested objects by clicking objects via a touch screen or a mouse when watching the video. Based on different attention tracking methods, the system architecture applies a corresponding attention data pre-processor to extract the attention objects.

[0035] For example, the gaze data pre-processor 114 obtains raw gaze data 140 from the eye-tracking module 106. The gaze data pre-processor 114 determines one or more eye tracking features based on processing (e.g., pre-processing) the raw gaze data 140. For example, the gaze data pre-processor 114 can determine one or more fixations and/or saccades from the raw gaze data 140. The raw gaze data 140 is associated with a sampling rate (e.g., 60 Hz). In some instances, the gaze data pre-processor 114 down-samples the raw gaze data 140 to map to the video frames from the video 104 if required. In some embodiments, the gaze data pre-processor 114 and/or the attention extractor 112 obtains user input. For example, the user can use the user input (e.g., alternative input) to indict objects that the user 102 is interested in by clicking on objects via a touch screen (e.g., a touch display screen), a mouse, and/or other input devices (e.g., keyboard). The gaze data pre-processor 114 and/or the attention extractor 112 obtains user input from the input devices indicating the objects of interest (e.g., the attention objects) and/or other types of user input.

[0036] Step 4. Object detection and registration 108: This module 108 firstly detects and registers all the objects in the video, then uniquely identifies each object presented in the input video frames 104. For example, object detection models such as you only look once (YOLO) or single shot detection can be used for this purpose. The output 142 (e.g., the annotated video frames) of this neural network are video frames with rectangles and their object identifiers (IDs).

[0037] For example, the object detection and registration module 108 (e.g., an object detection and registration device) uses one or more algorithms such as object detection models (e.g., YOLO, single shot detection, one or more neural networks, and/or other machine learning / artificial intelligence algorithms and/or models) to register (e.g., determine) objects within the video data 104. For instance, the object detection and registration module 108 obtains training data 144 from the training dataset 110 and uses the training data 114 to train the one or more algorithms. The object detection and registration module 108 uses the one or more algorithms (e.g., YOLO) to detect or register objects within the video data 104. For instance, using the one or more algorithms, the object detection and registration module 108 determines video frames with shapes (e.g., rectangles) and their object IDs (e.g., an object ID for each of the shapes). The object detection and registration module 108 provides the output 142 to the object attention extractor 116.

[0038] Step 5. Object attention extractor 116, 124: This module 116 enables the method to automatically identify the attention object in the beginning of the video based on a viewer's perspective by utilizing gaze data and describe the video completely from the object's perspective. In some instances, it is not required to watch full video to generate the report. The object attention extractor accepts annotated video frames (e.g., detected objects with their object IDs) and pre-processed gaze data from the gaze data preprocessor 114 as input. Firstly, this module 116 updates the information for each object in each frame after identifying the attention object based on fixation overlap. Every object is detected, for instance, has tags indicating whether it has attention or not, whether it is visible or not in the frame. For example, when the eye tracking module 106 detects a fixation, it sends the corresponding gaze point coordinates to the object attention extractor 116. Then, a check is performed of whether any of the objects within the frame intersect with the gaze point coordinates. If an object intersects with the gaze point, the object is tagged as having attention. If no objects intersect with the gaze point, then no object is tagged as having attention. The first step's output can be a dictionary with a frame as the key and a dictionary of objects as the value, which contains information on whether or not each object has attracted attention and is visible. An example of such a dictionary is shown in FIG. 2. FIG. 2 illustrates an example dictionary 200 for object attention extraction as an intermediate step according to an embodiment of the present invention. Then, this dictionary 200

is parsed to extract those objects that have continuous attention for at least an amount (e.g., 90%) of the fixation threshold (for instance, 10 frames). For example, the system can consider 90% of the fixation threshold to allow for human distractions. These are classified or tagged as attention objects that the viewer focused on in the video. In second phase, for each object, the frames in which this object is visible are extracted. This module can generate the frame sequences 124 corresponding to each attention object. For each focused object, the object attention extractor module 116 can output sequences of frames and the object feature vector. The object feature vector is a mathematical representation of an object's visual characteristics extracted from an image or video frame. The object feature vector includes a set of numerical values that encode information about the object's color, shape, texture, size, and other relevant features. These values can be used to compare objects, classify them into categories, or recognize them in new images or videos. Embodiments of the present invention use this feature vector to consider focused objects while generating description of the video.

[0039] For example, the object attention extractor 116 (e.g., the object attention extractor module or device) obtains the pre-processed gaze data from the gaze data preprocessor 114 and the annotated video frames (e.g., the detected objects with their object IDs) from the object detection and registration module 108. Using the pre-processed gaze data and the annotated video frames 142, the object attention extractor 116 determines one or more attention objects (e.g., objects, entities, people, or other elements of interest to the user 102) from the video frames. For instance, the object attention extractor 116 updates the information (e.g., meta-data) for each object in each frame from the annotated video frames 142 with the object IDs based on the fixation overlap. For example, for each detected object (e.g., each object detected by the module 108), the object includes tags (e.g., information and/or metadata) indicating whether the object has attention (e.g., the user 102 paid attention to it) or does not have attention (e.g., the user 102 did not pay enough attention to it). As mentioned above, fixation is the duration of an interval when a person's (e.g., user's 102) sight is fixed on a stationary object. Based on the fixation information and the object IDs from the annotated video frames 142, the object attention extractor 116 determines the tags (e.g., whether the object has attention or not) for the detected objects. The object attention extractor 116 outputs an object output (e.g., a dictionary) indicating the determined tags for the objects. For instance, FIG. 2 shows an example dictionary 200. The dictionary 200 includes a header, "object" 202. Then, the dictionary 200 includes two frames ("frame1" 204 and "frame2" 212), which indicate frames (e.g., annotated frames) of the video. Each frame includes detected objects such as objects 206-210 and 214-220. As such, certain objects (e.g., "object2" and "object3") are shown on both frames 204-212. The tags include whether the object 206-210 and 214-220 are visible in the frame (e.g., "is_visible")

and/or has the user's 102 attention (e.g., "has_attention") with a "0" indicating that the object is not visible or does not have the user's attention and "1" indicating the object is visible and has the user's 102 attention.

[0040] After determining (e.g., generating) the dictionary (e.g., dictionary 200), the object attention extractor 116 parses the dictionary to extract objects that have a continuous attention span of the user 102. For instance, the object attention extractor 116 compares the metadata (e.g., tags) associated with the objects from the dictionary with one or more threshold such as a fixation threshold (e.g., 10 frames) and an attention threshold (e.g., 90%). The attention threshold can allow for human distractions. Based on the comparison, the object attention extractor 116 determines the objects that have a continuous attention span as attention objects. For example, if the object attention extractor 116 determines that the user 102 focuses on an object for 10 frames with an attention threshold of 90%, then the object attention extractor 116 classifies the object as an attention object.

[0041] After classifying and/or tagging the objects as attention objects, the object attention extractor 116 extracts frames in which this object is visible. For instance, the object attention extractor 116 extracts one or more frames in which the object has an "is_visible" tag of "1". The object attention extractor 116 generates frame sequences 124 (e.g., frames of the interested objects such as the attention objects) corresponding to each attention object. Additionally, and/or alternatively, for each focused object (e.g., attention object), the object attention module 116 outputs the sequence of frames 124 and the object feature vector.

[0042] Step 6. Object-based video report generation module 120: This module 120 enables generation of the description of video for each attention object. Each report describes the context of a corresponding object across all related frames in the video. The module 120 receives the sequence of frames 124 for the attention objects and/or the object feature vector from the attention extractor 112 (e.g., the object attention extractor 116).

[0043] Step 6.1. Semantic segmentation module associated with a semantic segmentation model 126 (e.g., a semantic segmentation module 126): The semantic segmentation module 126 is used to understand the semantic of the image (e.g., what is in the image and where). This module 126 takes a sequence of frames extracted corresponding to each object (e.g., attention object) and object's feature vector 124 as input to a deep learning-based image segmentation model (e.g., a deep learning-based image semantic segmentation model) and generates labels for each frame. For example, if a frame contains a red car, a supermarket and a tree, then these can be labels for that particular frame. Any segmentation network such as U-Net, a fully-convolutional network, and/ or an encoder-decoder-based model can be used for this segmentation task. In some instances, this network can be trained, for example, on the

MICROSOFT Common Objects in Context (MS-COCO) dataset 122, which has 330,000 images (e.g., greater than 200,000 labelled) and 80 common object classes. The images in the dataset 122 are everyday objects captured from everyday scenes.

[0044] For example, the semantic segmentation module 126 (e.g., a semantic segmentation device 126) uses training data 156 from the training dataset 122 to train one or more models (e.g., a deep learning-based image segmentation model). For instance, the training dataset 122 can be the MS-COCO dataset 122, and the semantic segmentation module 126 uses the training data 156 within the MS-COCO dataset 122 for training the one or more models. Further, the semantic segmentation module 126 obtains the sequence of images (e.g., frames) and the object(s) 150 (e.g., the sequence of extracted frames and/or the object's feature vector 124). The semantic segmentation module 126 inputs the sequence of extracted frames and/or the object's feature vector 124 into the model (e.g., the deep learning-based image segmentation model) and generates one or more labels. For instance, the semantic segmentation module 126 generates labels for each frame such as a red car, a supermarket, and/or a tree for a particular frame.

[0045] Step 6.2. Video description module associated with a video description model 128 (e.g., a video description module 128): This module 128 describes the sequence of frames corresponding to each object. It can be any video summarizer trained on a domain-specific dataset 118 (e.g., CCTV footage video and summary), which takes a sequence of frames corresponding to each object and its feature vector as input and describe the frames with respect to this object in text. As the result, the video description module 128 generates individual textual reports for each object, and each report describes the context of the corresponding object across all related frames.

[0046] For example, the video description module 128 uses training data 146 from a training dataset 118 (e.g., a domain-specific dataset such as CCTV footage video and summary) to train a model (e.g., a video summarizer model, a neural network, and/or other types of machine learning / artificial intelligence models). The video description module 128 obtains the sequence of images (e.g., frames) and the object(s) 148 (e.g., the sequence of extracted frames and/or the object's feature vector 124). The video description module 128 inputs the sequence of extracted frames and/or the object's feature vector 124 into the model and outputs text describing the frames with respect to the attention object. As a result, the video description module 128 generates individual textual reports (e.g., the textual reports indicating the object ID and video description 130) for each object, and each report describes the context of the corresponding object across all related frames. The video description module 128 provides output 152 (e.g., text) indicating the object ID and video descriptions 130.

[0047] Step 6.3. Name entity recognizer 134 Generated reports (e.g., the reports 130) are then passed to the name entity recognizer 134. This module 134 identifies the named entities such as 'a red car' and 'a supermarket' from the text. It could be a neural network based language model, which is trained on an annotated training dataset 136. Additionally, and/or alternatively, the user 102 can provide a domain-specific dataset to fine-tune the name entity recognizer 134 for better performance. For instance, to fine-tune a named entity recognizer 134 for better performance, the user 102 can provide a domain-specific dataset including annotated examples of the entity types they are interested in. This dataset 136 can then be used to fine-tune the model's weights and biases to improve its accuracy for that domain. This process helps the model to learn domain-specific features, resulting in better performance on the specific task.

[0048] For instance, the name entity recognizer 134 trains a neural network based language model (e.g., a name entity recognition model) using training data 158 from a training dataset 136 (e.g., an annotated training dataset). Further, after the video description module 128 generates the reports 130, the name entity recognizer 134 obtains the reports (e.g., the text 154 from the reports 130). The name entity recognizer 134 uses the neural network based language model to identify the named entities within the text 154 from the reports 130.

[0049] Step 6.4. Entity consistency optimizer 132: After generating entity labels from both the semantic segmentation model 126 and the name entity recognizer 134, the system architecture 100 forwards both results to the entity consistency optimizer 132, which measures the consistency degree between entities found by the semantic segmentation model 126 from the object frames and entities found by the name entity recognizer 134 from the generated object description. This consistency degree is then added to the loss function of the video description model 128. For instance, the entity consistency optimizer 132 measures the consistency degree between two set of entities, and can calculate a similarity score between them. For example, the entity consistency optimizer 132 can use a contrastive loss to determine the degree of overlap between the two sets. The resulting loss can be interpreted as a measure of how consistent the two sets of entities are with each other. Once the consistency degree is calculated, the consistency degree is added to the loss function of the video description model 128. The loss function is a measure of how well the model 128 is able to predict the correct description for a given video. By incorporating the consistency degree into the loss function, the model 128 is encouraged to generate descriptions that are consistent with the entities present in the video. For instance, the semantic segmentation model 126 provides entities that are present in a frame such as red car, a tree, supermarket and name entity recognizer provides entities from the description like a car, supermarket. The loss value can be high and effectively penalizing the model 128 if it generates description that does not include all the entities.

[0050] In some embodiments, the video description model 128 aims to generate textual description, which covers as many entities as possible in the object frames by optimizing the loss function. As the result, the entity consistency optimizer 132 guides the video description model 128 to generate a comprehensive description without losing information. For instance, by incorporating the consistency degree into the loss function, the model 128 can be trained to generate more accurate and consistent descriptions of videos without losing important information. In practice, the entity consistency optimizer 132 can be a neural network model (e.g., an entity consistency optimizer neural network), for example a multi-layer perceptron (MLP), which computes the consistency score such as a contrastive loss between two entity results. The loss function is differentiable with respect to the MLP parameters, allowing it to be optimized using gradient-based methods such as backpropagation.

[0051] For instance, the entity consistency optimizer 132 obtains the results from the semantic segmentation model 126 and the name entity recognizer 134 (e.g., the identified name entities from the text 154 from the reports 130 and the labels from the semantic segmentation model). The entity consistency optimizer 132 determines (e.g., measures) a consistency degree between entities (e.g., labels) from the semantic segmentation model 126 and entities (e.g., the identified name entities from the text 154) from the name entity recognizer 134. The entity consistency optimizer 132 then adds the consistency degree to a loss function and provides the result (e.g., the loss function with the consistency degree) to the video description model 128. The video description model 128 then performs another iteration based on the loss function. As the result, the entity consistency optimizer 132 guides the video description model 128 to generate a comprehensive description without losing information. The entity consistency optimizer 132 can be and/or include a neural network model (e.g., MLP), which computes the consistency score (e.g., consistency degree) based on the entities from the semantic segmentation model 126 and the name entity recognizer 134.

[0052] Step 7. End the system 130, 138: After generating individual textual descriptions for each focused object as a list of <object ID, video description> 130 (e.g., the reports), the system architecture 100 sends them back to the user 102 and the process ends. Alternatively, as demonstrated in FIG. 1, the <object ID, video description> results 130 can be further forwarded to an optional decision module 138. Since a user 102 can have attention to multiple objects, the decision module 138 is designed to (e.g., configured) decide the best result, which means the most focused object from the list and returns a single report of the most focused object to the user 102. For instance, the decision module 138 can rank the objects based on their attention and take the top object as the final output. For example, the decision module 138 assigns a score to each attention object based on the attention duration. Then, the decision module 138 ranks the

attention objects in descending order based on their score, with the highest-scoring object being selected as the final output. For instance, in a CCTV video evidence and the user 102 places attention on two people (e.g., the offender and other person), the decision module 138 can assign a higher score to an offender because of having maximum attention than to the other person. This optional procedure is advantageous for the case where the user 102 is a system or part of another system, which directly consumes the output for their further operation.

[0053] For example, after one or more iterations of using the entity consistency optimizers 132 and the video description model 128, the object based video report generation module 120 outputs the reports 130 (e.g., a list of <object ID, video description>). For instance, the object based video report generation module 120 can output the reports 130 directly to the user 102 (e.g., display the reports on a computing device). The reports 130 can be displayed on the same display device as the display device that output the video data 104 or on a different display device. Additionally, and/or alternatively, the object based video report generation module 120 can output the reports 130 to the decision module 138. The decision module 138 is configured to determine the best result (e.g., the most focused object), and provide a single report to the user 102 (e.g., to the display device that displays the report to the user 102). For instance, the reports 130 can include a plurality of reports for a plurality of objects (e.g., each report is associated with an object and the video description for the object). The decision module 138 ranks and determines the top object, and provides the report associated with the top object to the user 102 (e.g., the display device that displays the report to the user 102).

[0054] FIG. 3 schematically illustrates an environment for eye-tracking-based object-focused video description according to an embodiment of the present invention. For example, the environment 300 shows example hardware (e.g., computing devices, computing systems, and databases) that are configured to perform one or more embodiments of the present invention (e.g., one or more embodiments of the present invention described in FIG. 1). The environment 300 is merely an example and other types environments are contemplated herein that can perform embodiments of the present invention. For instance, the environment 300 includes an eye tracking device 302 (e.g., eye tracking equipment and/or module 106) that is configured to track the eye movements of a user (e.g., user 102).

[0055] The environment 300 further includes a database 306 that includes the video data 308 (e.g., the video data 104) and/or training dataset(s) 310 (e.g., the training datasets 110, 118, 122, and/or 136). In some instances, the database 306 is a distributed database. For example, a first database can store the video data 308 and one or more additional databases can store the training dataset(s) 310.

[0056] The environment 300 includes a computing system 304. The computing system 304 includes one or more computing devices, computing platforms, systems, servers, desktops, laptops, tablets, mobile devices (e.g., smartphone device, or other mobile device), or any other type of computing device that generally comprises one or more communication components, one or more processing components, and one or more memory components. In some variations, the computing system 304 can be implemented as engines, software functions, and/or applications. In other words, the functionalities of the computing system 304 can be implemented as software instructions stored in storage (e.g., memory) and executed by one or more hardware processors.

[0057] The computing system 304 includes the attention extractor device 312 (e.g., the attention extractor 112), the object detection and registration device 314 (e.g., the object detection & registration 108), the object based video report generation device 316 (e.g., the object based video report generation module 120), and the decision device 318 (e.g., the decision module 138). The devices 312-318 can be hardware devices and/or processors (e.g., separate computing devices that are configured to perform the method described above). In some instances, the devices 312-318 are implemented as engines, software functions, and/or applications. In other words, the functionalities of the devices 312-318 can be implemented as software instructions stored in storage (e.g., memory) and executed by one or more hardware processors.

[0058] In an embodiment, the present invention can be applied to effect further improvements in the technical field of automated public safety and law enforcement tools, for example for improving the functionality of forensic tools. One use case can be for improving forensic tools for evidence reporting where video such as CCTV footage is a crucial evidence in a criminal investigation. Such footage allows investigators to watch the entire incident. The video contains all the information about sequence of events, criminal's activities and their entry and exit points. In short, these recordings can be used to prove or disprove allegations against the suspect. Whenever a case is summoned, CCTV evidence is comprehensively summarized on paragraph four of the MG5 report. It takes a lot of time for police officers to manually summarize the CCTV footage for what the offender did in that video. Embodiments of the present invention provide an automated computerized tool for police officers to have an efficient approach to help them fill out this report automatically based on a police officer's focused object (what they saw in that video), even for just a short part of the video. In this use case, the data source includes CCTV video frames and generated gaze data. Police officers can watch the video with an eye-tracking device to generate the gaze data. Application of the method according to an embodiment of the present invention firstly generates the raw gaze data and pre-processes it to calculate eye-tracking features like fixation and saccades, and to uniquely identify objects (e.g., a person) in the video frames. Then, using fixation overlap, the method identifies

the attention objects (e.g., offender and some other person or object) and generates the description of the CCTV footage with respect to the attention objects. For instance, the description is of what the offender did in that CCTV video. The system outputs a list of video descriptions specific to each attention object. For instance, if the police officer watched that video and gave attention to two persons (offender and victim), then the system can generate two reports corresponding to each attention object. As a resulting automated decision or action, physical change or technicity, the generated video description specific to the offender (the person that the police office first puts attention while watching) can be input to a computer-aided reporting system, which can utilize this report to automatically fill paragraph four of the MG5 report digitally.

[0059] Another use case is crime prevention or lead discovery, e.g. for drug dealer or terrorist tracking. In some cases, police officers want to go through extensive (very long, e.g., 10+ hours) CCTV video when looking into a case (such as one involving drug trafficking or terrorism) in order to find some leads for further investigation. It is very time consuming to watch the video to track the activities of a specific person (e.g., a drug dealer or terrorist suspect). Embodiments of the present invention provide an automated computerized tool for police officers to have an efficient approach to help them generate a report based on their focused object (what they saw in that video, even for few minutes). In this use case, the data source includes CCTV video frames and generated gaze data. Police officers can watch the video with an eye-tracking device to generate gaze data. Application of the method according to an embodiment of the present invention firstly generates the raw gaze data and pre-processes it to calculate eye-tracking features like fixation and saccades, and uniquely identifies objects (e.g. person) in the beginning from the video frames. Then, using fixation overlap, the system identifies the attention objects (e.g., drug dealer and some other person or object) and generates the description of the CCTV footage with respect to the attention objects. For instance, the description is of what activities the drug dealer did in that CCTV video. A list of video description specific to attention objects is output. For instance, if a police officer watched that video and gave attention to two persons (drug dealer and another person or object), then the system can generate two reports corresponding to each attention objects. As a resulting automated decision or action, physical change or technicity, the generated video description specific to the offender can be used to classify what this object did in that video (e.g., harm, no harm). If the object is a criminal, that rectangle box can be extracted from the frame and provided as input to an existing digital tracking system, and if CCTV footage identified the same object an alarm could be triggered automatically or automatic tracking could take place.

[0060] Embodiments of the present invention can also be practically applied to output video descriptions from the perspective of attention objects to effect further improvements in a number of other fields as well, for example to initiate automated decisions or actions based on the generated description. For example, such descriptions could be used by personalization systems for objects in videos that were the focus of particular users (e.g., to recommend certain products), for modifying videos to try to shift a viewer's focus, for training pilots, surgeons or other professions where object focus is important, among many other applications where video analysis or object focus plays a role.

[0061] In an embodiment, the present invention provides a method for automatically generating a video description from the perspective of an attention object, comprising the steps of:

Setup a new system (cold start):

1. Create or use an eye-tracking module 106, which generates the eye-tracking data (e.g., raw gaze data), and if a user provides a video, then watch the video using eye-tracking equipment associated with the system architecture 100.
2. Create or use a gaze data preprocessor module 114 to extract eye-tracking features like fixations and saccades.
3. Create an object detection and registration module 108 as follows:
    a. Train any object detection deep neural network or use a pre-trained model to identify the objects.
    b. For each frame, identify the objects as in step 3.a. and register them for unique identification.
4. Create or use an object attention extractor module 116 to identify the attention objects and their corresponding frames based on a fixation threshold.
5. Train an image segmentation model 126 on, e.g., the MS-COCO dataset 122, to identify entities shown in each frames.
6. Create a name entity recognizer 134, e.g., by training a neural network-based entity extractor, which explores the patterns of the important entities from the given documents to identify the important entities from the unknown document.
7. Create a video description module 128 by training any video summarization model on a domain-specific dataset (e.g., training dataset 118) to describe the sequences of frames corresponding to each object.
8. Create an entity consistency optimizer 132, e.g., MLP, by learning a similarity function between the two entities' sets.

When new data arrives (setup system in use):

1. The user inputs video data and watches the video using an eye-tracker associated with the system (eye-tracking module 106).

2. The system generates the reports for each object the user puts attention on by utilizing gaze data information.

a. Gaze data preprocessor module 114 pre-processes the raw gaze data and extracts the eye-tracking features.

b. Object detection and registration module 108 identifies the objects in the frames and registers them by giving them unique IDs.

c. Object attention extractor module 116 first identifies the attention objects based on fixation overlap and then extracts the frames corresponding to each attention object. This provides the enhanced functionality to automatically identify the attention object in the beginning of the video based on the viewer's perspective by utilizing gaze data and describe the video completely from object's perspective. It is not required to watch full video to generate the report.

d. Generation of individual textual reports for each attention object following two methods:

1. First method:
    a. Video description model (128) generates individual textual reports for each attention object, and each report describes the context of the corresponding object across all related frames.
2. Second method:
    a. Semantic segmentation module (126) generates the labels for each frame (for e.g., red car, tree etc.).
    b. Video description model 128 generates individual textual reports for each attention object, and each report describes the context of the corresponding object across all related frames (e.g., video descriptions of the attention object throughout the video data).

    c. Name entity recognizer 134 identifies the entities in reports generated.
    d. Entity consistency optimizer 132 compares entities from the name entity recognizer 134 with scene understanding results from the semantic segmentation model 126 to optimize the video description module 128. This provides the enhanced functionality that the system will not miss important information in the report. It utilizes the entity consistency optimizer 132 to compare the entities from name entity recognizer 134

with the entity labels generated from semantic segmentation module 126 to optimizes the quality of description generated by video descriptor.

[0062] Embodiments of the present invention provide for the following improvements over existing technology:

1. A method that automatically identifies the attention object in the beginning of the video based on a viewer's perspective by utilizing gaze data and describes the video completely from object's perspective. It is not required to watch full video to generate the report.

2. A method that generates a description of video for each attention object. Each report describes the context of corresponding object across all related frames in the video.

3. The system will not miss important information in the report. It utilizes an entity consistency optimizer to compare the entities from a name entity recognizer with entity labels generated from a semantic segmentation module using a machine learning model by learning a similarity function between them and then by incorporating the loss from the machine learning model into an overall loss of the video descriptor, and it optimizes the quality of description generated by video descriptor.

4. Extraction of attention object at the beginning without the need for watching the full video to generate an focused object-oriented video description. Xu, Jia, et al., "Gaze-enabled Egocentric Video Summarization via Constrained Submodular Maximization," IEEE Computer Society Conference on Computer Vision and Pattern Recognition:2235-2244 (2015), which is hereby incorporated by reference herein, uses a constrained submodular maximization approach to select the most relevant frames. The relevance score of each frame is calculated based on both visual and gaze-based features. In contrast, embodiments of the present invention first identify the attention object and then describe the video completely from this object's perspective using a self-optimized video descriptor, and it is not required to watch the full video to generate the report. Rather, if the viewer watches for just a few seconds, then the method according to an embodiment of the present invention can identify which object the viewer puts attention on and generate the report for that specific attention object automatically.

5. Ability to automatically recognize or identify entities. Ji, Zhong, et al., "Video Summarization with Attention-Based Encoder-Decoder Networks," arXiv:1708.09545v2 (2018), which is hereby incorporated by reference herein, uses a long short-term memory (LSTM)-based encoder to extract features from the input video frames and an attention-based LSTM-based decoder to generate a summary of the video. The attention mechanism allows the decoder to focus on the most relevant frames in the input video. On top of that, an embodiment of the present invention can be used to optimize the quality of the video descriptions by utilizing the name entity recognizer on the generated report. Firstly, all the entities present in each frame are

identified (e.g., red car, supermarket, tree etc.) using any segmentation model then the name entity recognition module is used to identify all the entities present in the generated video description. Lastly, the entity consistency optimizer 132 compares the entities from the name entity recognizer with the entity labels generated from the semantic segmentation module using the MLP by learning a similarity function between them and then by incorporating the loss from the MLP into the overall loss of video descriptor, and then it optimizes the quality of the description generated by video descriptor based in the identified entities.

[0063] FIG. 4 is a block diagram of an exemplary processing system, which can be configured to perform any and all operations disclosed herein. Referring to FIG. 4, a processing system 400 can include one or more processors 402, memory 404, one or more input/output devices 406, one or more sensors 408, one or more user interfaces 410, and one or more actuators 412. Processing system 400 can be representative of each computing system disclosed herein.

[0064] Processors 402 can include one or more distinct processors, each having one or more cores. Each of the distinct processors can have the same or different structure. Processors 402 can include one or more central processing units (CPUs), one or more graphics processing units (GPUs), circuitry (e.g., application specific integrated circuits (ASICs)), digital signal processors (DSPs), and the like. Processors 402 can be mounted to a common substrate or to multiple different substrates.

[0065] Processors 402 are configured to perform a certain function, method, or operation (e.g., are configured to provide for performance of a function, method, or operation) at least when one of the one or more of the distinct processors is capable of performing operations embodying the function, method, or operation. Processors 402 can perform operations embodying the function, method, or operation by, for example, executing code (e.g., interpreting scripts) stored on memory 404 and/or trafficking data through one or more ASICs. Processors 402, and thus processing system 400, can be configured to perform, automatically, any and all functions, methods, and operations disclosed herein. Therefore, processing system 400 can be configured to implement any of (e.g., all of) the protocols, devices, mechanisms, systems, and methods described herein.

[0066] For example, when the present disclosure states that a method or device performs task “X” (or that task “X” is performed), such a statement should be understood to disclose that processing system 400 can be configured to perform task “X”. Processing system 400 is configured to perform a function, method, or operation at least when processors 402 are configured to do the same.

[0067] Memory 404 can include volatile memory, non-volatile memory, and any other medium capable of storing data. Each of the volatile memory, non-volatile memory, and any other type of memory can include multiple different memory devices, located at multiple distinct locations and each having a different structure. Memory 404 can include remotely hosted (e.g., cloud) storage.

[0068] Examples of memory 404 include a non-transitory computer-readable media such as RAM, ROM, flash memory, EEPROM, any kind of optical storage disk such as a DVD, a Blu-Ray® disc, magnetic storage, holographic storage, a HDD, a SSD, any medium that can be used to store program code in the form of instructions or data structures, and the like. Any and all of the methods, functions, and operations described herein can be fully embodied in the form of tangible and/or non-transitory machine-readable code (e.g., interpretable scripts) saved in memory 404.

[0069] Input-output devices 406 can include any component for trafficking data such as ports, antennas (i.e., transceivers), printed conductive paths, and the like. Input-output devices 406 can enable wired communication via USB®, DisplayPort®, HDMI®, Ethernet, and the like. Input-output devices 406 can enable electronic, optical, magnetic, and holographic, communication with suitable memory 404. Input-output devices 406 can enable wireless communication via WiFi®, Bluetooth®, cellular (e.g., LTE®, CDMA®, GSM®, WiMax®, NFC®), GPS, and the like. Input-output devices 406 can include wired and/or wireless communication pathways.

[0070] Sensors 408 can capture physical measurements of environment and report the same to processors 402. User interface 410 can include displays, physical buttons, speakers, microphones, keyboards, and the like. Actuators 412 can enable processors 402 to control mechanical forces.

[0071] Processing system 400 can be distributed. For example, some components of processing system 400 can reside in a remote hosted network service (e.g., a cloud computing environment) while other components of processing system 400 can reside in a local computing system. Processing system 400 can have a modular design where certain modules include a plurality of the features/functions shown in FIG. 4. For example, I/O modules can include volatile memory and one or more processors. As another example, individual processor modules can include read-only-memory and/or local caches.

[0072] While embodiments of the invention have been illustrated and described in detail in the drawings and foregoing description, such illustration and description are to be considered illustrative or exemplary and not restrictive. It will be understood that changes and modifications may be made by those of ordinary skill within the scope of the following claims. In particular,

the present invention covers further embodiments with any combination of features from different embodiments described above and below. Additionally, statements made herein characterizing the invention refer to an embodiment of the invention and not necessarily all embodiments.

[0073] The terms used in the claims should be construed to have the broadest reasonable interpretation consistent with the foregoing description. For example, the use of the article "a" or "the" in introducing an element should not be interpreted as being exclusive of a plurality of elements. Likewise, the recitation of "or" should be interpreted as being inclusive, such that the recitation of "A or B" is not exclusive of "A and B," unless it is clear from the context or the foregoing description that only one of A and B is intended. Further, the recitation of "at least one of A, B and C" should be interpreted as one or more of a group of elements consisting of A, B and C, and should not be interpreted as requiring at least one of each of the listed elements A, B and C, regardless of whether A, B and C are related as categories or otherwise. Moreover, the recitation of "A, B and/or C" or "at least one of A, B or C" should be interpreted as including any singular entity from the listed elements, e.g., A, any subset from the listed elements, e.g., A and B, or the entire list of elements A, B and C.

CLAIMS

What is claimed is:

1. A method for generating attention based video description using eye-tracking, comprising:

obtaining, using an eye-tracker device, raw gaze data associated with a user watching at least a portion of video data comprising a plurality of frames;

identifying, based on the raw gaze data, attention objects associated with objects of interest within the video data and extracting one or more frames from the plurality of frames based on the identified attention objects;

generating one or more individual textual reports for each of the identified attention objects based on the one or more extracted frames; and

outputting the one or more individual textual reports that describe the video data in context of each of the identified attention objects.

2. The method of claim 1, further comprising:

receiving user input from one or more input devices indicating user indicated interested objects from within the video data, wherein identifying the attention objects is based on the user input.

3. The method of claim 1 or 2, further comprising:

pre-processing the raw gaze data to extract one or more eye tracking features associated with the user watching the video data, wherein the one or more eye tracking features comprise one or more fixations, wherein the one or more fixations indicate one or more durations of when the user's sight is fixed on one or more stationary objects from the video data; and

using one or more object detection models and the video data to generate annotated video frames indicating video objects and unique identifiers associated with the video objects,

wherein identifying the attention objects is based on the one or more extracted eye tracking features and the annotated video frames.

4. The method of claim 3, wherein identifying the attention objects comprises:

determining one or more tags for each of the video objects based on fixation overlap between the one or more extracted eye tracking features and the annotated video frames, wherein the one or more tags indicate whether the user paid attention to the video objects while watching the video data;

generating a dictionary based on the one or more tags for the video objects;

comparing the one or more tags for each of the video objects with a fixation threshold and an attention threshold; and

classifying one or more video objects of the video objects as the attention objects based on the comparison.

5. The method of any of the preceding claims, wherein generating the one or more individual textual reports for each of the identified attention objects is further based on optimizing generated descriptions for the one or more individual textual reports by utilizing a name entity recognition model and a deep learning-based image semantic segmentation model.

6. The method of claim 5, wherein generating one or more individual textual reports for each of the identified attention objects comprises:

generating one or more first individual textual reports for each of the identified attention objects based on the one or more extracted frames; and

generating one or more second individual textual reports for each of the identified attention objects based on the first individual textual reports and a consistency degree, wherein the one or more second individual textual reports indicate one or more object identifiers associated with the identified attention objects and video descriptions of the identified attention object throughout the video data, and wherein outputting the one or more individual textual reports comprises outputting the one or more second individual textual reports.

7. The method of claim 6, wherein generating the one or more second individual textual reports for each of the identified attention objects comprises:

inputting the one or more extracted frames associated with the identified attention objects into the deep learning-based image semantic segmentation model to generate labels for the one or more extracted frames;

identifying named entities from the one or more first individual textual reports using the name entity recognition model;

generating the consistency degree based on the named entities, the generated labels, and an entity consistency optimizer neural network; and

generating the one or more second individual textual reports based on optimizing a video description model using the consistency degree and a loss function associated with the video description model.

8. The method of claims 6 or 7, wherein the one or more second individual textual reports comprises a plurality of second individual textual reports for a plurality of attention objects, wherein each of the plurality of second individual textual reports is associated with a single attention object, of the plurality of attention objects, and a video description for the single attention object, wherein outputting the one or more second individual textual reports comprises:

ranking the plurality of attention objects; and

providing for display a top ranked attention object from the plurality of attention objects.

9. The method of any of claims 3-8, wherein obtaining the raw gaze data comprises generating, using an eye-tracking module for the eye-tracker device, the raw gaze data based on the user watching the video data, and

wherein pre-processing the raw gaze data comprises extracting, using a gaze data preprocessor module, the one or more eye tracking features, wherein the one or more eye tracking features comprise the one or more fixations and one or more saccades, wherein the one or more saccades indicate rapid eye movement of the user when shifting focus from a first stationary object to a second stationary object.

10. The method of any of claims 3-9, wherein the one or more object detection models is stored in an object detection and registration module, wherein the one or more object detection models is an object detection deep neural network that is trained based on using training data or a pre-trained model that identifies the video objects, wherein the one or more object detection models identifies the video objects in each video frame of the plurality of frames and registers the video objects for unique identification.

11. The method of any of claims 4-10, wherein identifying the attention objects is based on using an object attention attractor module, and wherein determining the attention objects based on the dictionary comprises determining the attention objects within their corresponding frames of the video data based on the dictionary and a fixation threshold.

12. The method of any of claims 5-11, further comprising:

training the deep learning-based image semantic segmentation model to identify labels within frames of a video using one or more segmentation training datasets, wherein the one or more segmentation training datasets comprises a MICROSOFT Common Objects in Context (MS-COCO) dataset; and

training a neural network based entity extractor to generate the name entity recognition model, wherein the neural network based entity extractor determines patterns of entities from given documents to identify important entities within the given documents.

13. The method of any of the preceding claims, further comprising:

training a video summarization model using a domain specific dataset to generate a video description module, wherein the video summarization model describes sequences of frames corresponding to objects within the sequence of frames; and

learning a similarity function between two entity sets using an entity consistency optimizer, wherein the entity consistency optimizer is a Multi-Player Perception (MLP) neural network.

14. A system for generating attention based video description using eye-tracking, the system comprising one or more hardware processors, which, alone or in combination, are configured to provide for execution of the following steps:

obtaining, using an eye-tracker device, raw gaze data associated with a user watching video data comprising a plurality of frames;

identifying, based on the raw gaze data, attention objects associated with objects of interest within the video data and extracting one or more frames from the plurality of frames based on the identified attention objects;

generating one or more individual textual reports for each of the identified attention objects based on the one or more extracted frames; and

outputting the one or more individual textual reports that describe the video data in context of each of the identified attention objects.

15. A tangible, non-transitory computer-readable medium having instructions thereon which, upon being executed by one or more processors, alone or in combination, provide for execution of a method for generating attention based video description using eye-tracking comprising the following steps:

obtaining, using an eye-tracker device, raw gaze data associated with a user watching video data comprising a plurality of frames;

identifying, based on the raw gaze data, attention objects associated with objects of interest within the video data and extracting one or more frames from the plurality of frames based on the identified attention objects;

generating one or more individual textual reports for each of the identified attention objects based on the one or more extracted frames; and

outputting the one or more individual textual reports that describe the video data in context of each of the identified attention objects.

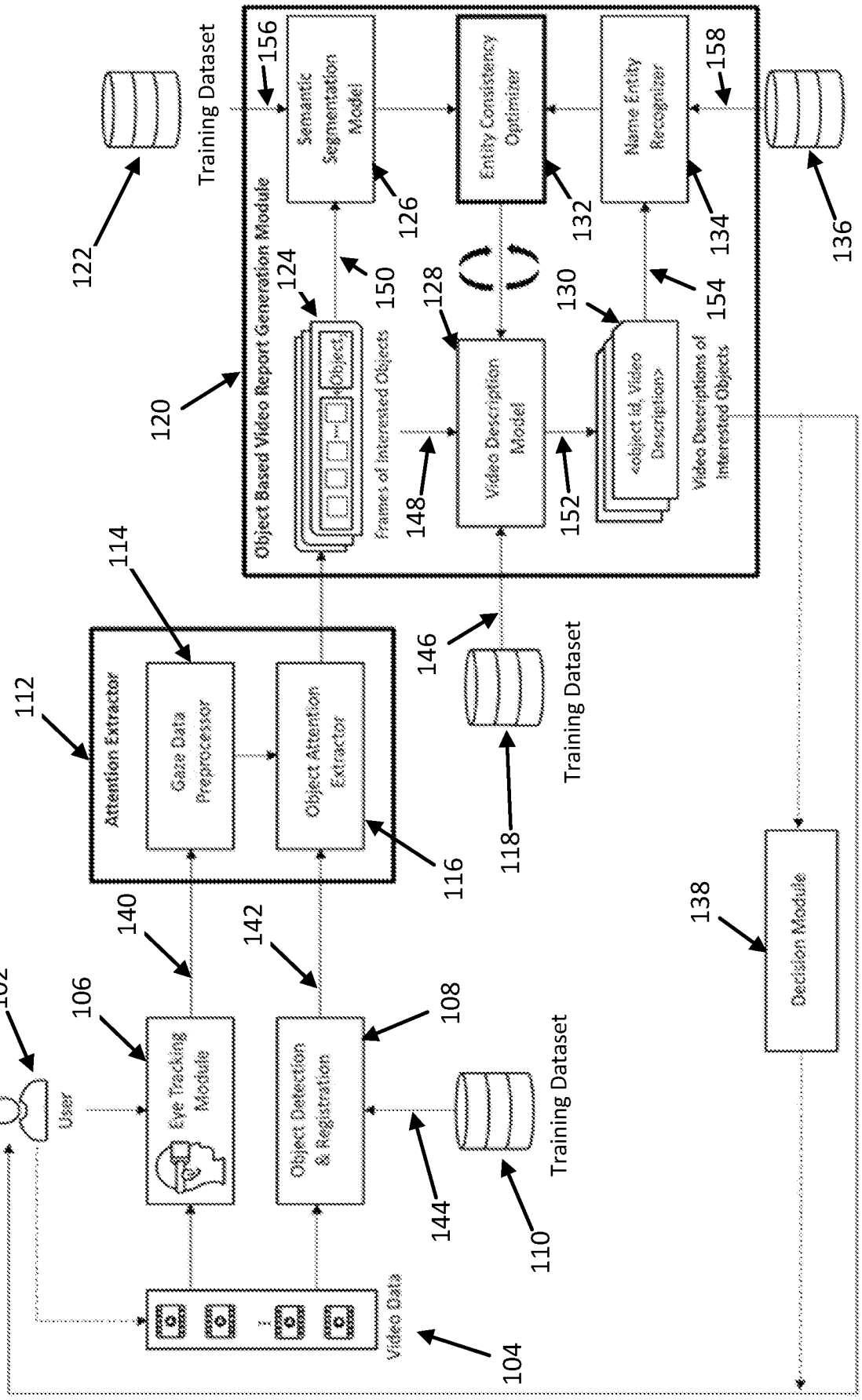
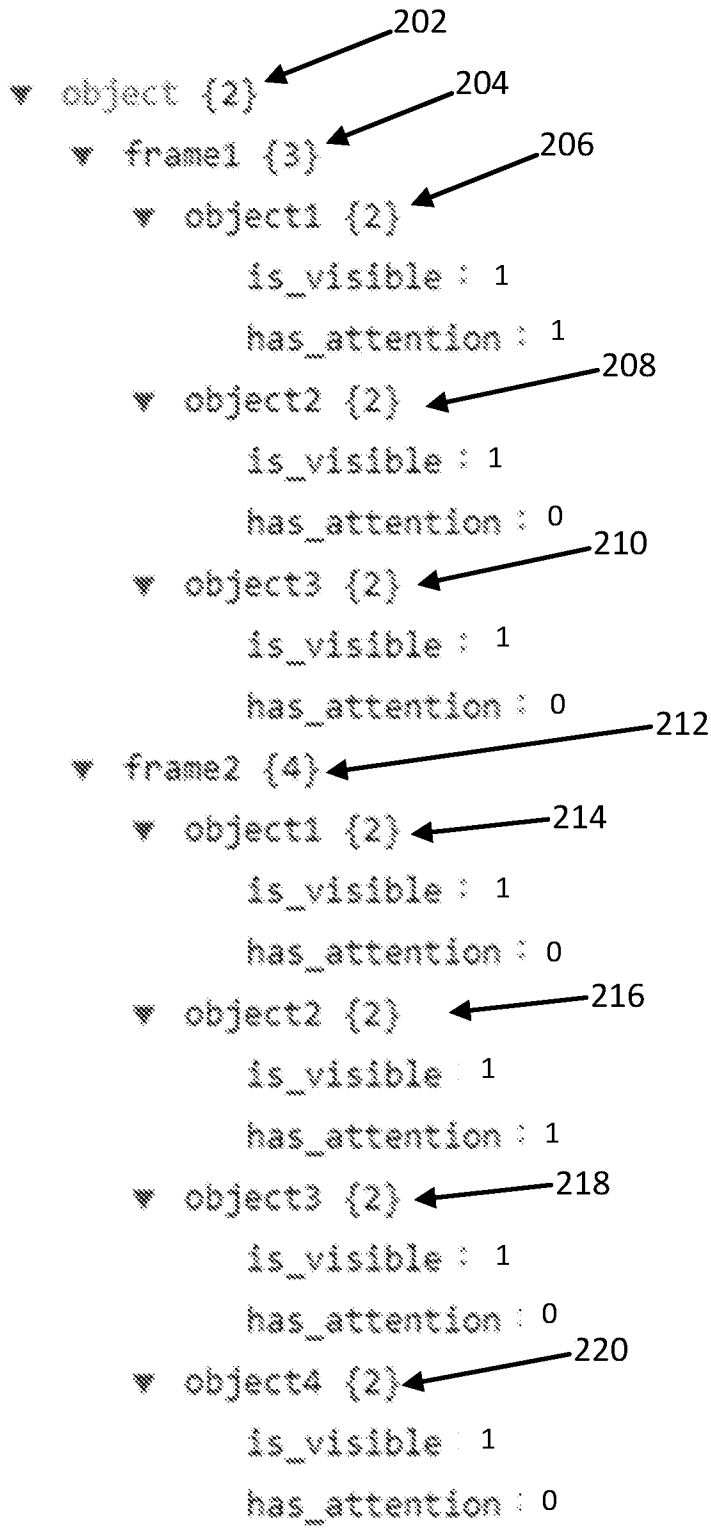
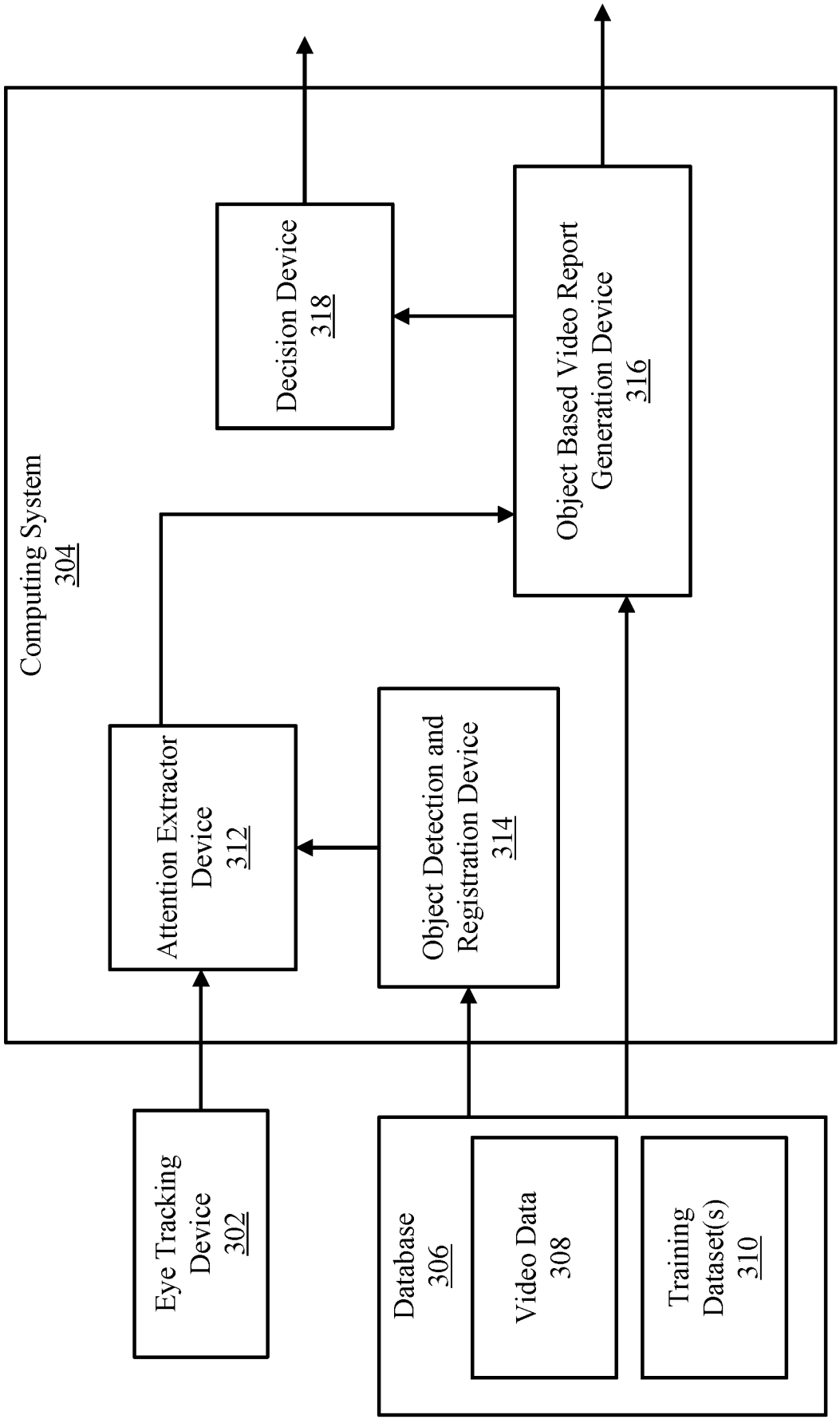

Training Dataset
122
Object Based Video Report Generation Module
120
156
Semantic Segmentation Model
126
Entity Consistency Optimizer
132
Name Entity Recognizer
158
134
Training Dataset
136
124
Object
Frames of Interested Objects
150
128
Video Description Model
148
130
<object id, Video Description>
Video Descriptions of Interested Objects
154
152
114
Attention Extractor
112
Gaze Data Preprocessor
Object Attention Extractor
116
146
Training Dataset
118
140
142
102
User
106
Eye Tracking Module
Object Detection & Registration
108
138
Decision Module
144
110
Training Dataset
Video Data
104
100

FIG. 1


202
object {2}
204
frame1 {3}
206
object1 {2}
is_visible : 1
has_attention : 1
208
object2 {2}
is_visible : 1
has_attention : 0
210
object3 {2}
is_visible : 1
has_attention : 0
212
frame2 {4}
214
object1 {2}
is_visible : 1
has_attention : 0
216
object2 {2}
is_visible : 1
has_attention : 1
218
object3 {2}
is_visible : 1
has_attention : 0
220
object4 {2}
is_visible : 1
has_attention : 0

200

FIG. 2


Computing System
304
Eye Tracking Device
302
Attention Extractor Device
312
Object Detection and Registration Device
314
Database
306
Video Data
308
Training Dataset(s)
310
Decision Device
318
Object Based Video Report Generation Device
316

300

FIG. 3

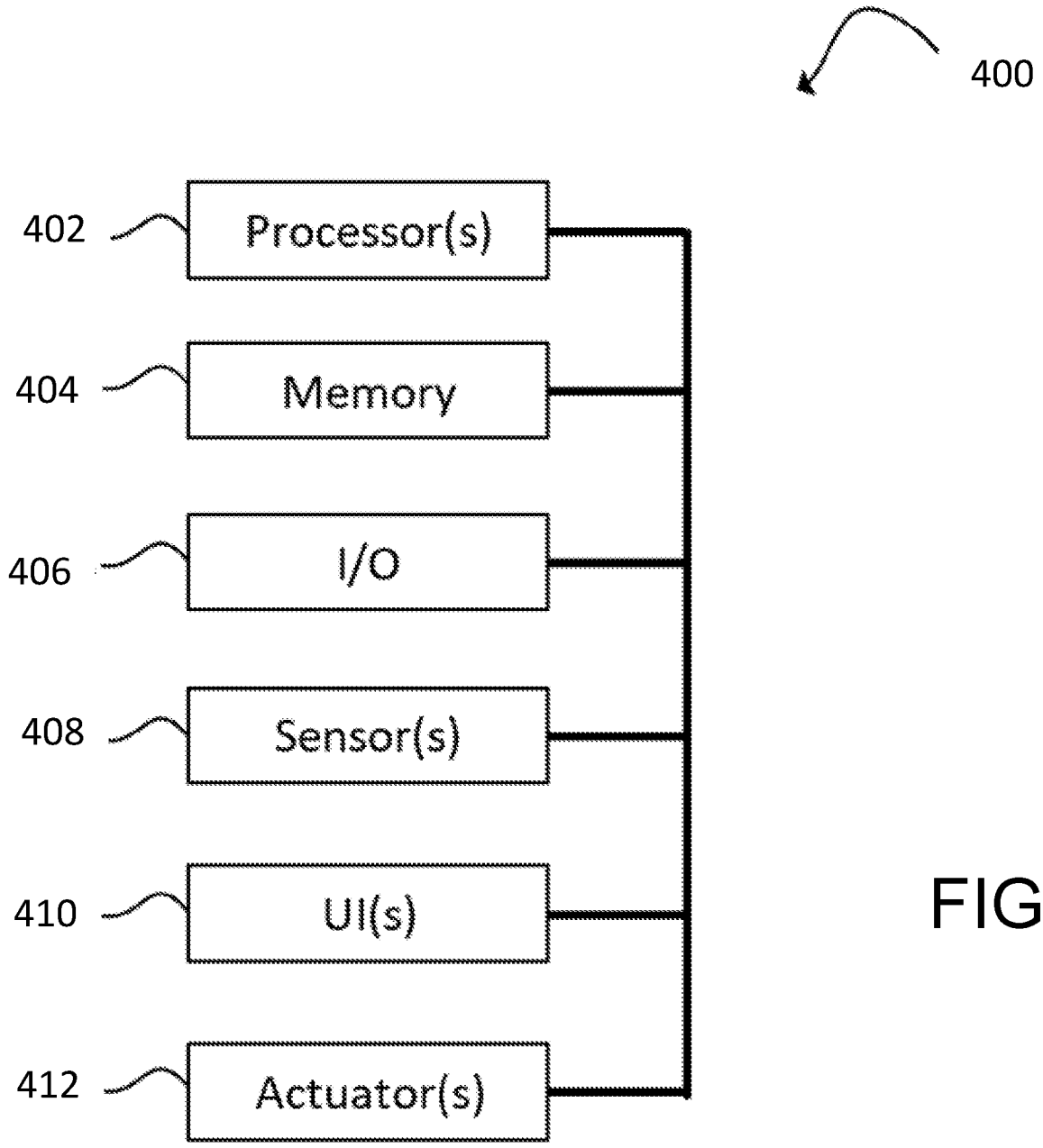

400
402
Processor(s)
404
Memory
406
I/O
408
Sensor(s)
410
UI(s)
412
Actuator(s)

FIG. 4

INTERNATIONAL SEARCH REPORT

International application No
PCT/IB2023/056504

A. CLASSIFICATION OF SUBJECT MATTER
INV. H04N21/234 G06V10/70 G06V20/40
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)
H04N G06V

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WU JIAXIN ET AL: "Gaze Aware Deep Learning Model for Video Summarization", 19 September 2018 (2018-09-19), SAT 2015 18TH INTERNATIONAL CONFERENCE, AUSTIN, TX, USA, SEPTEMBER 24-27, 2015; [LECTURE NOTES IN COMPUTER SCIENCE; LECT.NOTES COMPUTER], SPRINGER, BERLIN, HEIDELBERG, PAGE(S) 285 - 295, XP047486286, ISBN: 978-3-540-74549-5 [retrieved on 2018-09-19] pages 287-289; figure 1 ----- -/--	1-15

[x] Further documents are listed in the continuation of Box C. [ ] 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
8 November 2023	20/11/2023

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	Folea, Octavian

1

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

INTERNATIONAL SEARCH REPORT

International application No
PCT/IB2023/056504

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	XU JIA ET AL: "Gaze-enabled egocentric video summarization via constrained submodular maximization", 2015 IEEE CONFERENCE ON COMPUTER VISION AND PATTERN RECOGNITION (CVPR), IEEE, 7 June 2015 (2015-06-07), pages 2235-2244, XP032793666, DOI: 10.1109/CVPR.2015.7298836 [retrieved on 2015-10-14] pages 2236-223; figure 1 -----	1-15
X	WU JIAXIN ET AL: "Foveated convolutional neural networks for video summarization", MULTIMEDIA TOOLS AND APPLICATIONS, KLUWER ACADEMIC PUBLISHERS, BOSTON, US, vol. 77, no. 22, 30 April 2018 (2018-04-30), pages 29245-29267, XP036609532, ISSN: 1380-7501, DOI: 10.1007/S11042-018-5953-1 [retrieved on 2018-04-30] page 29248 - page 29253; figure 1 -----	1-15

1