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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
12 July 2018 (12.07.2018)

WIPO | PCT

(10) International Publication Number
WO 2018/127627 A1

(51) International Patent Classification:
G06F 17/30 (2006.01) *G06K 9/00* (2006.01)
H04N 21/8549 (2011.01) *G06N 3/04* (2006.01)
H04N 21/8405 (2011.01)

(21) International Application Number:
PCT/FI2018/050001

(22) International Filing Date:
02 January 2018 (02.01.2018)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
1700265.0 06 January 2017 (06.01.2017) GB

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

(72) Inventor: **CRICRI, Francesco**; Massunkuja 1 A 14, 33100 Tampere (FI).

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

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

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

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

(54) Title: METHOD AND APPARATUS FOR AUTOMATIC VIDEO SUMMARISATION

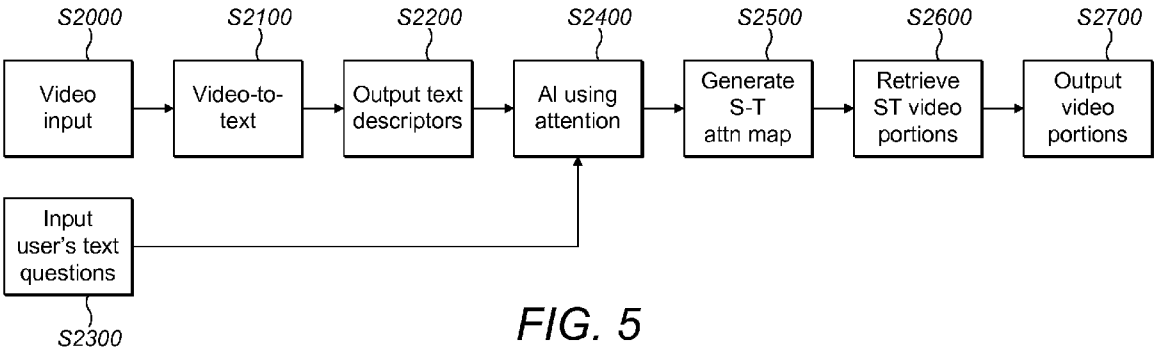

FIG. 5

(57) Abstract: This specification describes a method comprising: analysing (S2400), using a neural network, a text description of an input video and an input question; causing production of an attention map (S2500) based on the text description and the input question; and determining locations of the attention map having the highest attention value.

WO 2018/127627 A1

Method and Apparatus for Automatic Video Summarisation

Field

This specification generally relates to automatic video summarisation.

Background

Video summarisation includes producing a video which is smaller in size. Temporal video summarisation includes producing a shorter video. Spatial video summarisation includes producing a video which has less spatial extent that the original. Video summarisation may include detecting events in the video which are relatively more interesting than other events in the video.

Summary

According to a first aspect, the specification describes a method comprising: analysing, using a neural network, a text description of an input video and an input question; causing production of an attention map based on the text description and the input question; and determining locations of the attention map having the highest attention value.

The attention map may be a temporal attention map, wherein the locations correspond to temporal locations of the attention map having the highest attention value.

The attention map may be a spatial attention map, wherein the locations correspond to spatial locations of the attention map having the highest attention value.

The attention map may be a spatio-temporal attention map, wherein the locations correspond to spatial and temporal locations of the attention map having the highest attention value.

The method may further comprise outputting a video summary having video portions corresponding to the locations of the attention map having the highest attention value.

The method may further comprise selecting a video portion based on the temporal location of the attention map having the highest attention value and surrounding temporal locations having an attention value above a threshold attention value.

The method may further comprise converting the input video to the text description.

The method may further comprise converting the text description and input question respectively to a text description summary vector and a question summary vector.

The method may further comprise providing the text description summary vector and the question summary vector to the neural network.

According to a second aspect, the specification describes a computer program comprising machine readable instructions that, when executed by computing apparatus, causes it to perform any method as described with reference to the first aspect.

According to a third aspect, the specification describes an apparatus configured to perform any method as described with reference to the first aspect.

According to a fourth aspect, the specification describes an apparatus comprising: at least one processor; and at least one memory including computer program code which, when executed by the at least one processor, causes the apparatus to perform a method comprising: analysing, using a neural network, a text description of an input video and an input question; causing production of an attention map based on the text description and the input question; and determining locations of the attention map having the highest attention value.

The attention map may be a temporal attention map, wherein the locations correspond to temporal locations of the attention map having the highest attention value.

The attention map may be a spatial attention map, wherein the locations correspond to spatial locations of the attention map having the highest attention value.

The attention map may be a spatio-temporal attention map, wherein the locations correspond to spatial and temporal locations of the attention map having the highest attention value.

The computer program code, when executed, may cause the apparatus to perform: outputting a video summary having video portions corresponding to the locations of the attention map having the highest attention value.

The computer program code, when executed, may cause the apparatus to perform: selecting a video portion based on the temporal location of the attention map having the

highest attention value and surrounding temporal locations having an attention value above a threshold attention value.

The computer program code, when executed, may cause the apparatus to perform: converting the input video to the text description.

The computer program code, when executed, may cause the apparatus to perform: converting the text description and input question respectively to a text description summary vector and a question summary vector.

The computer program code, when executed, may cause the apparatus to perform: providing the text description summary vector and the question summary vector to the neural network.

According to a fifth aspect, the specification describes a computer-readable medium having computer-readable code stored thereon, the computer-readable code, when executed by at least one processor, causes performance of at least: analysing, using a neural network, a text description of an input video and an input question; causing production of an attention map based on the text description and the input question; and determining locations of the attention map having the highest attention value.

According to a sixth aspect, there is provided an apparatus comprising means for: analysing, using a neural network, a text description of an input video and an input question; causing production of an attention map based on the text description and the input question; and determining locations of the attention map having the highest attention value.

Brief Description of the Figures

For a more complete understanding of the methods, apparatuses and computer-readable instructions described herein, reference is now made to the following descriptions taken in connection with the accompanying drawings in which:

Figure 1 is a schematic illustration of an automatic video summariser, according to embodiments of this specification;

Figure 2 is a schematic illustration of temporal video summarisation according to embodiments of this specification;

Figure 3 is a schematic illustration of spatial video summarisation according to embodiments of this specification;

Figure 4 is a flow chart illustrating various operations which may be performed by the automatic video summariser in order to convert video to a text description according to embodiments of this specification;

Figure 5 is a flow chart illustrating various operations which may be performed by the automatic video summariser in order to produce a video summary based on a user's question according to embodiments of this specification;

Figure 6 is a flow chart illustrating operations which may be performed by the automatic video summariser in order to produce a spatio-temporal attention map according to embodiments of this specification;

Figure 7 illustrates an example of a spatio-temporal attention map produced by the automatic video summariser according to embodiments of this specification;

Figure 8 is a schematic illustration of an example configuration of the automatic video summariser according to embodiments of this specification;

Figure 9 is a computer-readable memory medium upon which computer-readable code may be stored, according to embodiments of this specification.

Detailed Description

In the description and drawings, like reference numerals may refer to like elements throughout.

Figure 1 is a schematic illustration of an automatic video summariser 10. The automatic video summariser 10 described herein make use of neural networks in order to produce spatio-temporal summaries including visual information relevant to a user's question or request. In this way, the events in the video which are considered to be relevant to the user's question are determined and video portions showing these events can be output as a spatio-temporal summary for the user.

The automatic video summariser 10 comprises a video-to-text module 20, an artificial intelligence (AI) attention module 30, a user interface 40 for receiving a user input, and an output 50, which may be a display, for example. The AI attention module may use deep learning methods such as attention mechanisms, neural attention mechanisms, or one or more neural networks outputting attention weights.

Figure 2 is a schematic illustration of temporal video summarisation. In temporal summarisation, the size of an input video 100 made up of video frames 100a-100i is reduced in size in terms of content by producing a video summary with a shorter time duration. A number of frames may be extracted from a video 100 formed of frames 100a-

i. For example, frames 100a,b,e,f,g,i may be extracted and joined temporally one after the other, maintaining the temporal order intact. The output video summary would comprise video portion 101 made up of frames 100a-b, video portion 102 made up of frames 100e-f, and video portion 103 made up of frames 100h-i. Accordingly, the summary will be a video having fewer frames than the input video. The portions may be made up of any number of frames. The portions may contain different frame numbers to the other portions. The temporal portions may be determined based on events occurring in the video. For example, a temporal portion may relate to one specific event occurring in the video. Selection of the temporal portions of the video may be performed as described in more detail with reference to Figures 4 to 7.

The video 100 may be a virtual reality video, for example a 360 degree video shot by a camera having a 360 degree field of view, such as the Nokia OZO camera. An example of a frame 110 from a virtual reality video can be seen in figure 3. The video may include multiple events in different spatial sectors of the video. The video may therefore be spatially summarised. A spatial video summary is a video comprising video crops, i.e. spatial video portions extracted from the original video by cropping spatially. Figure 3 illustrates spatial crops 111, 112, and 113. In spatial summarisation, the size of the video crops may be the same for all crops. In embodiments where the crops are not the same size, a resizing step may be applied to increase the resolution of at least one video crop. Increasing the resolution may be performed, for example, by upsampling with or without interpolation. Increasing the resolution may also be performed by using neural super-resolution methods. Alternatively, the resizing step may involve decreasing the resolution of at least one video crop. Decreasing the resolution may be performed, for example, by down-sampling of the video crop. Selection of the spatial portions of the video may be performed as described in more detail with reference to Figures 4 to 7.

By performing both temporal and spatial summarisation, a spatio-temporal video summary can be produced. For example, the video 100 may be a full length 360 degree movie. The movie may include multiple events temporally and multiple events spatially.

Figure 4 is a flow chart illustrating various operations which may be performed by the automatic video summariser in order to convert video to a text description. In some embodiments, not all of the illustrated operations need to be performed. Operations may also be performed in a different order compared to the order presented in Figure 4.

In operation S1000 the automatic video summariser may receive an input video from a video source. The video may be a video extract, or it may be a full length movie. The video may be provided from any suitable video source. For example, the video may be stored on a storage medium such as a DVD, Blue-Ray, hard drive, or any other suitable storage medium. Alternatively, the video may be obtained via streaming or download from an external server.

In operation S1100 an input video is analysed by a feature extraction module. The feature extraction module may comprise a Convolutional Neural Network (CNN). A CNN is an artificial neural network which represents currently the state-of-the-art for performing feature extraction from images and videos. A CNN consists of a sequence of computation layers, where the input is the data (a video frame or an image) and the output is a feature vector, i.e., a vector describing the input image. There may be different types of computation layers in a CNN, but the most important is the convolutional layer. A convolutional layer performs a convolution operation on its input, but using a set of convolution kernels. Other types of computation layers present in a CNN may be pooling layers, non-linear activation function layers, batch-normalization layers, etc. However, the present invention is not limited to a CNN and other feature extraction methodologies may be utilized.

In operation S1200, the features extracted in operation S1000 may be input to a temporal neural network. The temporal neural network may comprise a Recurrent Neural Network (RNN). A suitable RNN may be, for example, a Long Short-Term Memory network (LSTM).

In operation S1300, the temporal network outputs a “frame-description” vector, for each input video-frame. The frame description vector corresponds to a description of the video-frame. The frame-description vector may be used for generating a sentence or phrase describing the video frame, represented by a vector of real numbers.

In operation S1400, the frame description vectors may be analysed by a second RNN. The second RNN may also be a LSTM network, or any other suitable temporal neural network.

The second RNN generates a set of characters, or words, describing the input video-frame. As such a vector comprising a set of sentences describing the whole video is output.

In operation S1500, a softmax function is applied to the vector output by the second RNN as a result of operation S1400. This indicates the distribution of the words corresponding to the extracted features throughout the video. The vector which is output may be referred to as a "text description vector".

In operation S1600, an index synchronisation is performed. In order to determine the temporal locations of the features within the video, the text description is synchronised with the video. This includes associating each word or character with a certain video frame. A word or character may be associated with several adjacent video frames.

The association of the words or characters with corresponding video frames can be achieved by outputting a video-frame index for each word or character, corresponding to the index of the frame which is described by those words or characters. For example, in one case, one word may be associated with multiple adjacent frames.

In operation S1700, the automatic video summariser outputs a text description of the video associated with corresponding time indexes.

However, it will be recognised that any suitable implementation of a video to text module 20 can be utilised.

Figure 5 is a flow chart illustrating various operations which may be performed by the automatic video summariser in order to produce a spatio-temporal summary of an input video. In some embodiments, not all of the illustrated operations need to be performed. Operations may also be performed in a different order compared to the order presented in Figure 5.

In operation S2000 an input video is received.

In operation S2100, the video is converted to text, for example as described with reference to Figure 4. However, it will be understood that any suitable video to text conversion may be used.

In operation S2200, the automatic video summariser outputs text descriptions of the video.

In operation S2300 the automatic video summariser receives a user question or request. The question or request is input, or converted into, a text format. The question or request may relate to information the user would like to know about the input video. For example, the user may wish to find out whether there are any car crashes in the video. Therefore, the user may input a question such as "was there any car crashes in this movie?", or a request such as "would you summarise all the romantic scenes from the movie". The interface may be configured such that the user can input the question or request through user interface 40, for example by typing on a keyboard or on a touchscreen device connected to the automatic video summariser. Alternatively, the question or request may be verbally output by a user and received by voice recognition software to convert the question into text.

In operation S2400, the text question (or request) and the text descriptions of the video are input into an artificial intelligence (AI) attention module 30, which may comprise one or more neural networks, for example attention neural networks, and/or other operations which produce an "attention vector" . The text question and text description are analysed by the AI attention module. The question may be analysed before being input into the AI attention module 30. An example of how the question may be analysed is described in more detail with reference to Figure 6.

In operation S2500, the AI attention module 30 produces a spatio-temporal attention map representing the attention-intensity that a neural network has put at that point in time and spatial region when trying to answer the user's questions.

In step S2600, the automatic video summariser retrieves the spatial and temporal portions of the input video corresponding to the temporal and spatial locations of the spatio-temporal attention map having the highest attention-intensity values.

In step S2700, the automatic video summariser outputs the selected video portions as a spatio-temporal video summary.

Figure 6 is a flow chart illustrating in more detail the steps involved in producing the spatio-temporal attention map used in order to produce the spatio-temporal video summarisation. In some embodiments, not all of the illustrated operations need to be performed. Operations may also be performed in a different order compared to the order presented in Figure 6.

In operation S3000, the text descriptions output as a result of operation S1700 of Figure 4 are input to a word-embedding module.

In operation S3100, the word-embedding module converts the text descriptions to a set of dense vectors. Each of the dense vectors may represent a single word with a plurality of real numbers. The words in the text description are each converted from a vocabulary representation to a vector of real numbers. The vector of real numbers may be of lower dimensionality than the input vector of vocabulary entries, for example a vector with less dimensions or axes. The new representation is a point in an "embedding space", where words with similar semantics are nearby. The word-embedding module may be implemented by a multi-layer perceptron network or alternatively a single fully-connected layer. In general, the word-embedding module may transform an input into a more convenient output representation. For example, words may be transformed into a new representation for which similar words lie close to each other in the new representation space.

In operation S3200, the text description vectors are input to an RNN where the vectors are analysed. The RNN outputs a single output vector, which will be referred to herein as a text description summary vector. The RNN may be an LSTM.

In operation S3300, the question is input to a word-embedding module.

In operation S3400, the word-embedding module converts the question to a set of dense vectors. The words in the question are each converted from a vocabulary representation to a vector of real numbers with lower dimensionality, in a similar way to the text descriptions in operation S3100.

In operation S3500, the question vectors are input into an RNN where the vectors are analysed. The RNN outputs a single output vector which summarises the question, which will be referred to herein as a question summary vector. The RNN may be an LSTM.

In operation S3600, the text description summary vector and question summary vector are combined. The combination operation may be a concatenation in one of the dimensions of the input vectors, or an element-wise addition (if the input vectors have same dimensionalities). However, any suitable combination operation may be used at this step.

In operation S3700, the concatenated summary vectors are provided to a multi-layer perceptron (MLP) neural network. The MLP neural network may be referred to as an "attention neural network". The MLP is a neural network comprising a set of dense (i.e. fully connected) layers, followed by a softmax layer.

The dense layers of the MLP learn how to map the concatenated word-embedded text descriptions and user questions to an attention vector. The mapping is learned from data via a training process which happens offline, and which happens end-to-end for the whole model proposed in this invention. The input data is videos and a set of questions for each video, and the ground-truth output is the video segments which form the target video summary. The attention vector is in practice a set of attention weights (i.e., real numbers), summing up to 1, where each attention weight is associated to a certain temporal location of the video.

The softmax layer will output a probability distribution over "temporal attention weights" w.

The size of the output vector (i.e. the number of weights w) is the number of temporal locations, which is the number of words in the text describing the input video. In an alternative implementation, the size of the output vector is less than the number of words in the video description, and thus an attention weight can refer to more than one word. This would be a case where the attention is "quantized".

The weights represent a 1-dimensional "temporal attention map" (TAM), having bins which each correspond to a temporal location, and having a value of the value of the attention weight associated to that temporal location. The TAM value at location t, TAM[t], represents the attention-intensity that the attention neural network has put at that point in time when trying to answer the user's question.

The temporal locations associated with each bin correspond to temporal location of the input video. The attention weights output by the MLP are a vector of N bins, where N is the number of total temporal locations of the video. Therefore, the attention weights correspond to words of the text description and are arranged in the same temporal order as the temporal order of the words of the text description of the video. Accordingly temporal synchronisation is achieved based on the temporal location of the attention weights and the corresponding words of the text description. The dimensionality of the

vector output by the MLP is determined automatically based on the number of words of the text description created by the video-to-text module 20.

In operation S3800, the attention neural network outputs the probability distribution over attention weights which can be represented as the temporal attention map. A temporal location t^* of the TAM corresponding to the highest attention value in the TAM indicates the temporal location of the video which answers the user's question.

The temporal extent of the video portion is determined based on the temporal extent of the attention values around t^*. For example, a threshold value of the attention weight values may determine the temporal boundaries of the video portion to extract. That is, the video portion is selected based on temporal locations of attention weights above a given threshold. However, the temporal extent of the video portion may be selected in any other suitable way.

Figure 7 illustrates an example of a spatio-temporal attention map (STAM) produced by the automatic video summariser. The STAM represents the attention weights corresponding to each temporal location and spatial region of the video.

The TAM is extended to the spatial domain by analysing the video separately in the spatial dimension. For example, the video may be divided into a given number of angular sectors. Each sector is analysed separately by several attention networks. The joint output of the attention network is a 2-dimensional attention map, or "spatio-temporal attention map" (STAM).

The STAM is output as a matrix indexed using two indices, one for the time (t), and one for the space (the angular sector s). In Figure 7, the time runs along the x-axis of the map, and the space runs along the y-axis. In order to answer the user's question, the video portion (i.e. the particular temporal location and extent, and the spatial crop) will be determined by the highest value of attention within the STAM matrix. The video portion is based on the temporal location t^* and angular sector s^* having the highest attention value in the STAM matrix.

The video may be divided into a number of predetermined sectors. Alternatively, the division of the video may be dynamic. For example, the automatic video summariser may divide the video by means of "spatial scene cut detection". Spatial scene cut detection may be achieved by analysing the video with deep learning or multimedia analysis techniques

to detect objects, actions and activities, and then virtually cutting the scene to include the object, action and activity spatially. Therefore, the amount of data needed to analyse and summarize a spatial virtual reality video may be reduced.

Spatial summary may be applicable to 360 degree videos in order to convert a 360 degree video to a standard size video. Spatial summary may also be performed without any temporal summarisation, if this is desired.

In Figure 7, the temporal and spatial locations determined by the automatic video summariser as answering the user's question are indicated by the indices t1, s1, and t2, s2.

Therefore, the indices may be used to extract the corresponding temporal and spatial portions of the video for output to a user as a spatio-temporal summary of the video, based on the user's question. In this case, the user is provided with two video portions. The first video portion corresponds to the temporal location of the video indicated by indices t1, s1. The temporal extent of the video portion may be determined as described above, for example by setting a threshold attention value for the values temporally adjacent to t1. The second video portion corresponds to the temporal location of the video indicated by indices t2, s2.

Accordingly, by determining the highest attention values, the automatic video summariser is able to output a video summary which is determined to be the most relevant to the user's question or request. The video portions may be output through the output 50. The output may be a display which forms part of the automatic video summariser 10. Alternatively, the automatic video summariser 10 may be configured to output the video portions to a display which does not form part of the automatic video summariser 10, such as a display of a TV or PC, etc. For example, the automatic video summariser may be located on a server which is separate to the display through which the video portions are output. The automatic video summariser may be configured to output indicators of temporal locations of a video to be played in a video summary.

Figure 8 is a schematic block diagram of an example configuration of an automatic video summariser such as that described with reference to Figures 1 to 7. The video summariser may comprise memory and processing circuitry. The memory 11 may comprise any combination of different types of memory. In the example of Figure 8, the memory comprises one or more read-only memory (ROM) media 13 and one or more random

access memory (RAM) memory media 12. The processing circuitry 14 may be configured to process an input video and user question as described with reference to Figures 1 to 7.

The memory described with reference to Figure 8 may have computer readable instructions stored thereon 13A, which when executed by the processing circuitry 14 causes the processing circuitry 14 to cause performance of various ones of the operations described above. The processing circuitry 14 described above with reference to Figure 8 may be of any suitable composition and may include one or more processors 14A of any suitable type or suitable combination of types. For example, the processing circuitry 14 may be a programmable processor that interprets computer program instructions and processes data. The processing circuitry 14 may include plural programmable processors. Alternatively, the processing circuitry 14 may be, for example, programmable hardware with embedded firmware. The processing circuitry 14 may be termed processing means. The processing circuitry 14 may alternatively or additionally include one or more Application Specific Integrated Circuits (ASICs). In some instances, processing circuitry 14 may be referred to as computing apparatus.

The processing circuitry 14 described with reference to Figure 8 is coupled to the memory 11 (or one or more storage devices) and is operable to read/write data to/from the memory. The memory may comprise a single memory unit or a plurality of memory units 13 upon which the computer readable instructions 13A (or code) is stored. For example, the memory 11 may comprise both volatile memory 12 and non-volatile memory 13. For example, the computer readable instructions 13A may be stored in the non-volatile memory 13 and may be executed by the processing circuitry 14 using the volatile memory 12 for temporary storage of data or data and instructions. Examples of volatile memory include RAM, DRAM, and SDRAM etc. Examples of non-volatile memory include ROM, PROM, EEPROM, flash memory, optical storage, magnetic storage, etc. The memories 11 in general may be referred to as non-transitory computer readable memory media.

The term 'memory', in addition to covering memory comprising both non-volatile memory and volatile memory, may also cover one or more volatile memories only, one or more non-volatile memories only, or one or more volatile memories and one or more non-volatile memories.

The computer readable instructions 13A described herein with reference to Figure 8 may be pre-programmed into the automatic video summariser. Alternatively, the computer readable instructions 13A may arrive at the automatic video summariser via an

electromagnetic carrier signal or may be copied from a physical entity such as a computer program product, a memory device or a record medium such as a CD-ROM or DVD. The computer readable instructions 13A may provide the logic and routines that enable the automatic video summariser to perform the functionalities described above. For example, the video-text module 20, AI attention module 30, the feature extraction module, and the word-embedding module may be implemented as computer readable instructions stored on one or more memories, which, when executed by the processor circuitry, cause processing input data according to embodiments of the invention. The combination of computer-readable instructions stored on memory (of any of the types described above) may be referred to as a computer program or a computer program product.

Figure 9 illustrates an example of a computer-readable medium 16 with computer-readable instructions (code) stored thereon. The computer-readable instructions (code), when executed by a processor, may cause any one of or any combination of the operations described above to be performed.

As will be appreciated, the automatic video summariser described herein may include various hardware components which have may not been shown in the Figures since they may not have direct interaction with the shown features.

Embodiments may be implemented in software, hardware, application logic or a combination of software, hardware and application logic. The software, application logic and/or hardware may reside on memory, or any computer media. In an example embodiment, the application logic, software or an instruction set is maintained on any one of various conventional computer-readable media. In the context of this document, a "memory" or "computer-readable medium" may be any non-transitory media or means that can contain, store, communicate, propagate or transport the instructions for use by or in connection with an instruction execution system, apparatus, or device, such as a computer.

Reference to, where relevant, "computer-readable storage medium", "computer program product", "tangibly embodied computer program" etc., or a "processor" or "processing circuitry" etc. should be understood to encompass not only computers having differing architectures such as single/multi-processor architectures and sequencers/parallel architectures, but also specialised circuits such as field programmable gate arrays FPGA, application specific circuits ASIC, signal processing devices and other devices. References to computer program, instructions, code etc. should be understood to express software for

a programmable processor firmware such as the programmable content of a hardware device as instructions for a processor or configured or configuration settings for a fixed function device, gate array, programmable logic device, etc.

As used in this application, the term 'circuitry' refers to all of the following: (a) hardware-only circuit implementations (such as implementations in only analogue and/or digital circuitry) and (b) to combinations of circuits and software (and/or firmware), such as (as applicable): (i) to a combination of processor(s) or (ii) to portions of processor(s)/software (including digital signal processor(s)), software, and memory(ies) that work together to cause an apparatus, such as a mobile device or server, to perform various functions) and (c) to circuits, such as a microprocessor(s) or a portion of a microprocessor(s), that require software or firmware for operation, even if the software or firmware is not physically present.

This definition of 'circuitry' applies to all uses of this term in this application, including in any claims. As a further example, as used in this application, the term "circuitry" would also cover an implementation of merely a processor (or multiple processors) or portion of a processor and its (or their) accompanying software and/or firmware. The term "circuitry" would also cover, for example and if applicable to the particular claim element, a baseband integrated circuit or applications processor integrated circuit for a mobile phone or a similar integrated circuit in server, a cellular network device, or other network device.

If desired, the different functions discussed herein may be performed in a different order and/or concurrently with each other. Furthermore, if desired, one or more of the above-described functions may be optional or may be combined. Similarly, it will also be appreciated that the flow diagrams of Figures 4 to 6 are examples only and that various operations depicted therein may be omitted, reordered and/or combined.

Although various aspects are set out in the independent claims, other aspects comprise other combinations of features from the described embodiments and/or the dependent claims with the features of the independent claims, and not solely the combinations explicitly set out in the claims.

It is also noted herein that while the above describes various examples, these descriptions should not be viewed in a limiting sense. Rather, there are several variations and

modifications which may be made without departing from the scope of the appended claims.

Claims

1. A method comprising:

analysing, using a neural network, a text description of an input video and an input question;

causing production of an attention map based on the text description and the input question; and

determining locations of the attention map having the highest attention value.

2. A method according to claim 1, wherein

the attention map is a temporal attention map, and wherein the locations correspond to temporal locations of the attention map having the highest attention value.

3. A method according to claim 1, wherein the attention map is a spatial attention map, and wherein the locations correspond to spatial locations of the attention map having the highest attention value.

4. A method according to claim 1, wherein the attention map is a spatio-temporal attention map, wherein the locations correspond to spatial and temporal locations of the attention map having the highest attention value.

5. A method according to any preceding claim, comprising outputting a video summary having video portions corresponding to the locations of the attention map having the highest attention value.

6. A method according to claim 5, comprising selecting a video portion based on the temporal location of the attention map having the highest attention value and surrounding temporal locations having an attention value above a threshold attention value.

7. A method according to any preceding claim, further comprising converting the input video to the text description.

8. A method according to any preceding claim, further comprising converting the text description and input question respectively to a text description summary vector and a question summary vector.

9. A method according to claim 8, further comprising providing the text description summary vector and the question summary vector to the neural network.

10. A computer program comprising machine readable instructions that, when executed by computing apparatus, causes it to perform the method of any preceding claim.

11. Apparatus configured to perform the method of any of claims 1 to 9.

12. Apparatus comprising:

at least one processor; and

at least one memory including computer program code which, when executed by the at least one processor, causes the apparatus to perform a method comprising:

analysing, using a neural network, a text description of an input video and an input question;

causing production of an attention map based on the text description and the input question; and

determining locations of the attention map having the highest attention value.

13. A computer-readable medium having computer-readable code stored thereon, the computer-readable code, when executed by the at least one processor, causes performance of at least:

analysing, using a neural network, a text description of an input video and an input question;

causing production of an attention map based on the text description and the input question; and

determining locations of the attention map having the highest attention value.

14. Apparatus comprising means for:

analysing, using a neural network, a text description of an input video and an input question;

causing production of an attention map based on the text description and the input question; and

determining locations of the attention map having the highest attention value.

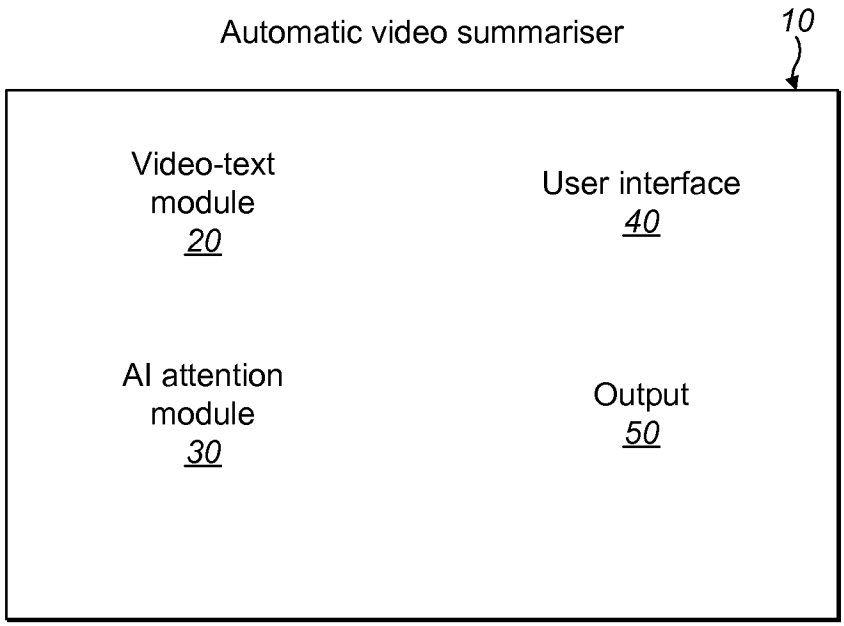
Automatic video summariser
10
Video-text module 20
User interface 40
AI attention module 30
Output 50

FIG. 1

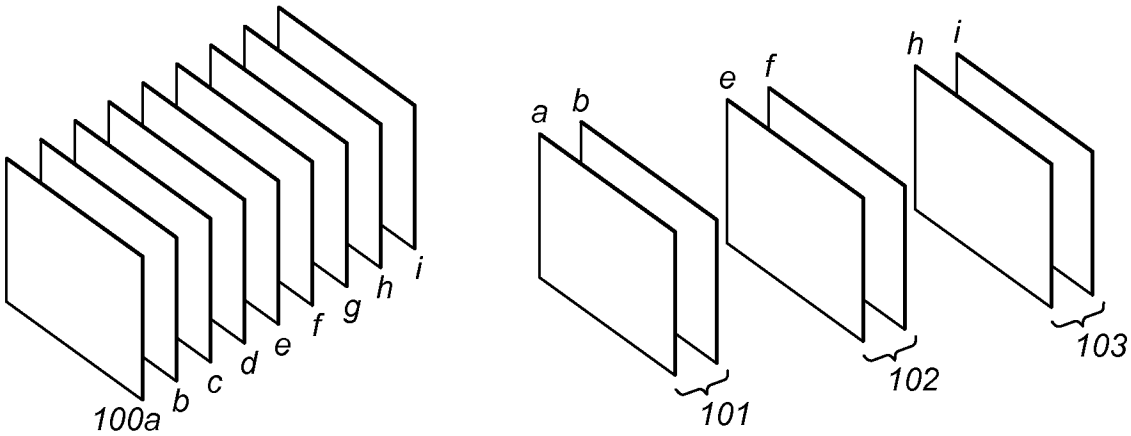

FIG. 2

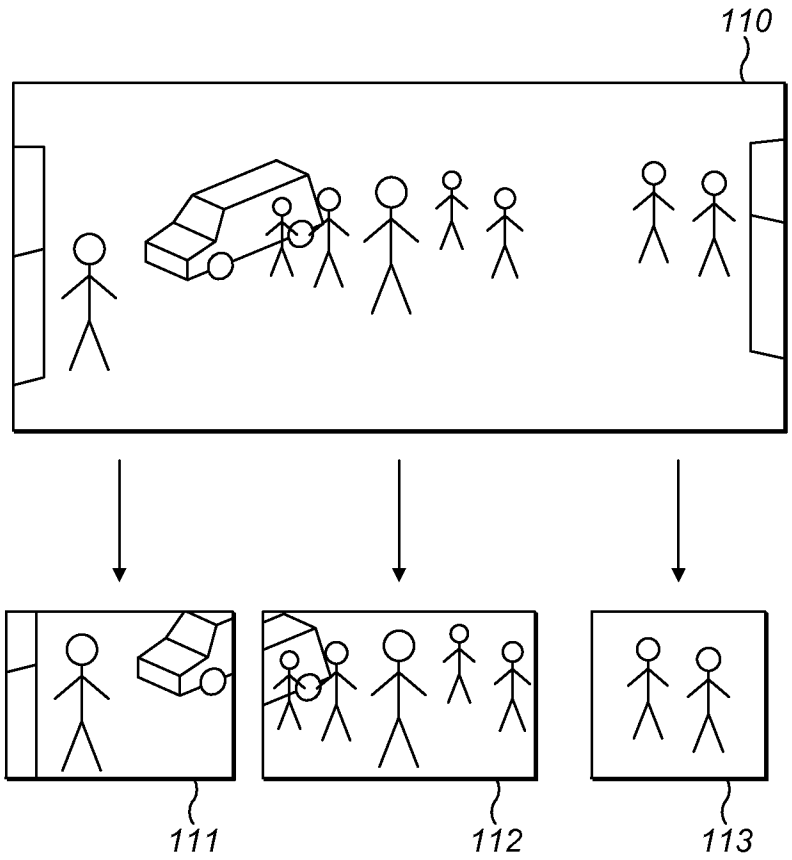

FIG. 3

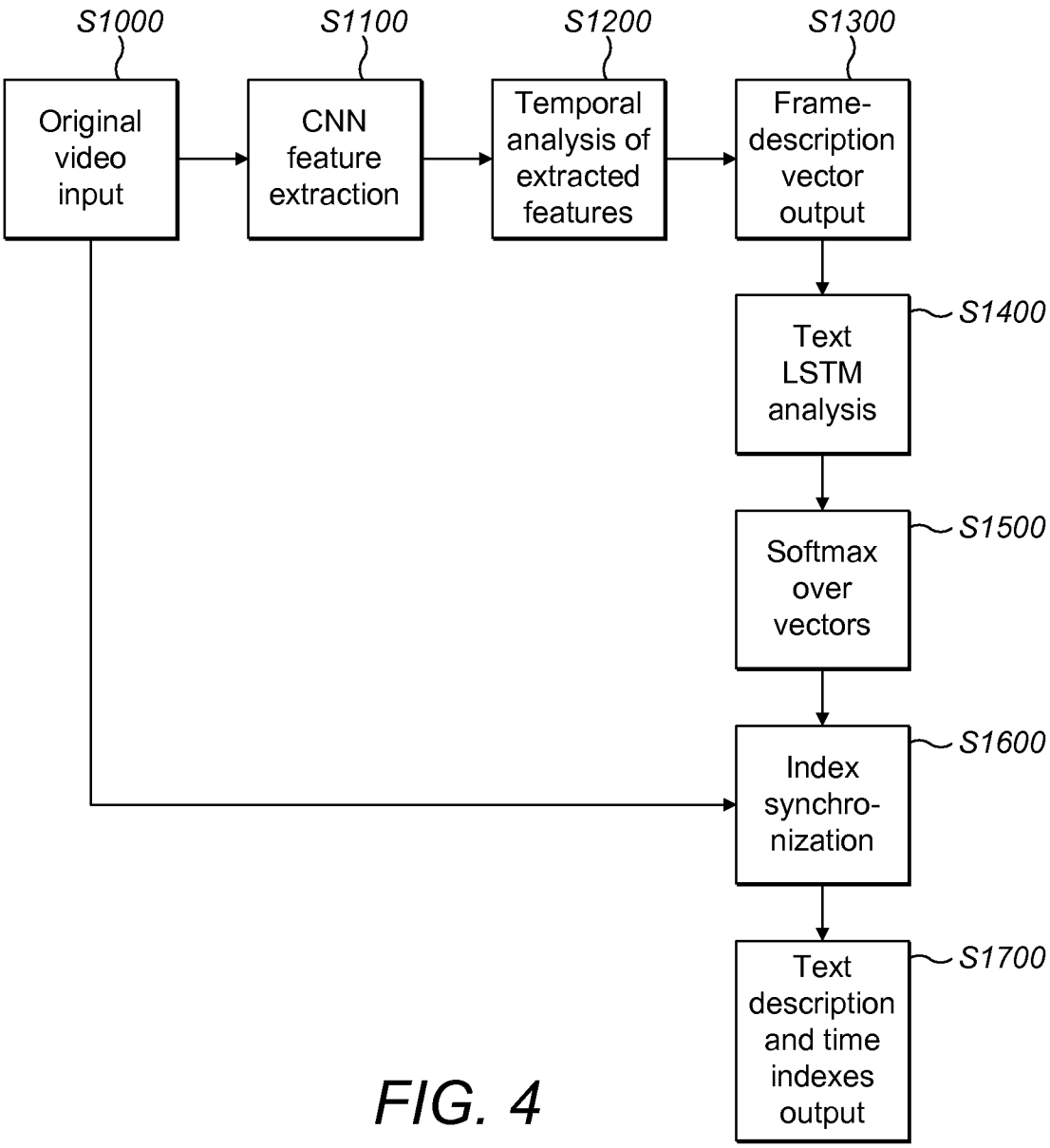
S1000
S1100
S1200
S1300
Original video input
CNN feature extraction
Temporal analysis of extracted features
Frame-description vector output
Text LSTM analysis
S1400
Softmax over vectors
S1500
Index synchro-nization
S1600
Text description and time indexes output
S1700

FIG. 4

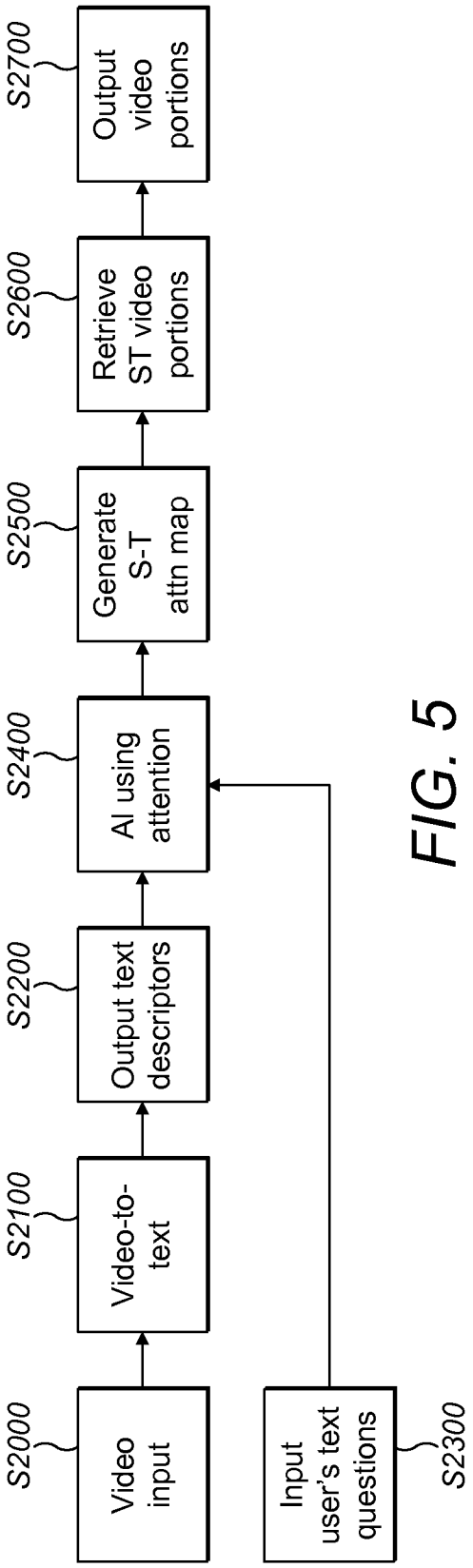
S2000
Video input
S2100
Video-to-text
S2200
Output text descriptors
S2400
AI using attention
S2500
Generate S-T attn map
S2600
Retrieve ST video portions
S2700
Output video portions
Input user's text questions
S2300

FIG. 5

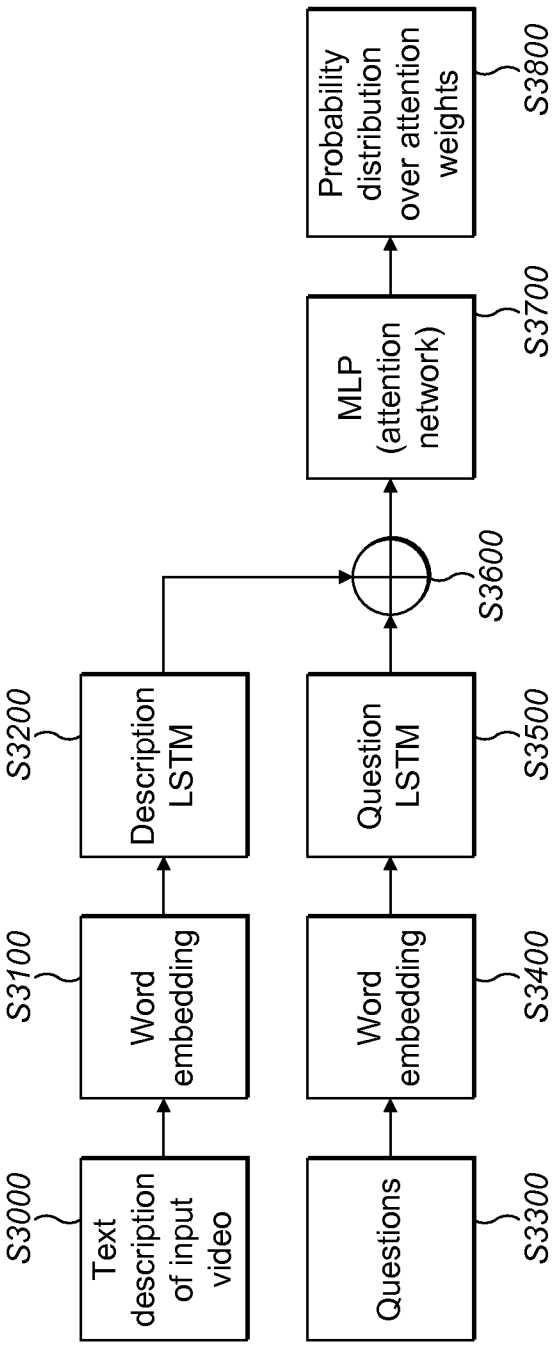
S3000
Text description of input video
S3100
Word embedding
S3200
Description LSTM
Questions
S3300
Word embedding
S3400
Question LSTM
S3500
S3600
MLP (attention network)
S3700
Probability distribution over attention weights
S3800

FIG. 6

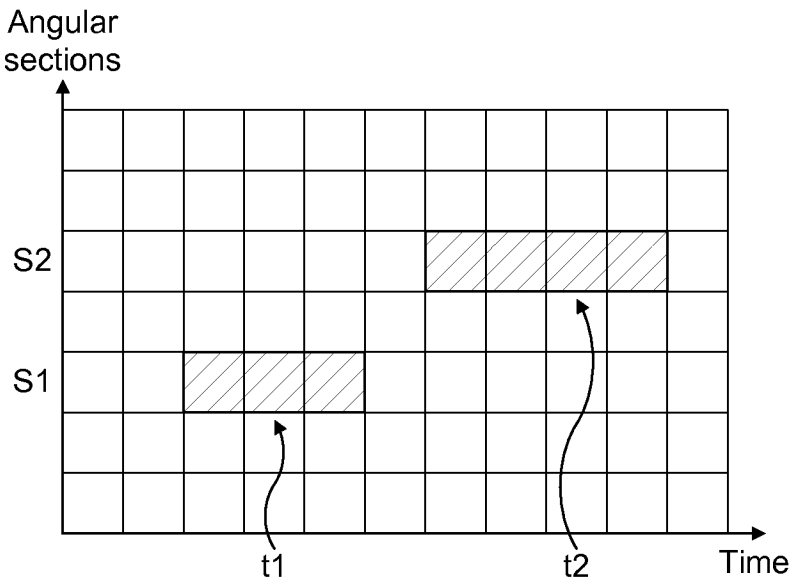
Angular sections
S2
S1
t1
t2
Time

FIG. 7

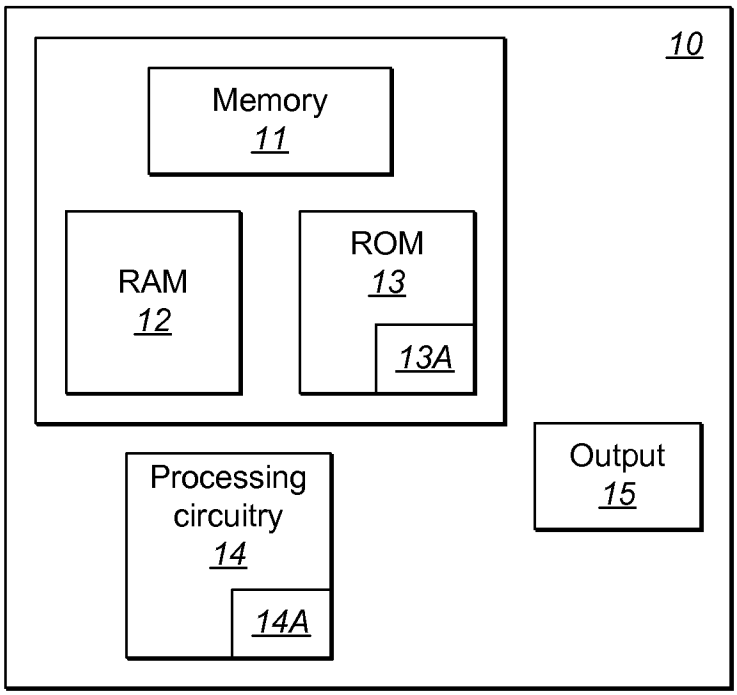
10
Memory
11
RAM
12
ROM
13
13A
Output
15
Processing circuitry
14
14A

FIG. 8

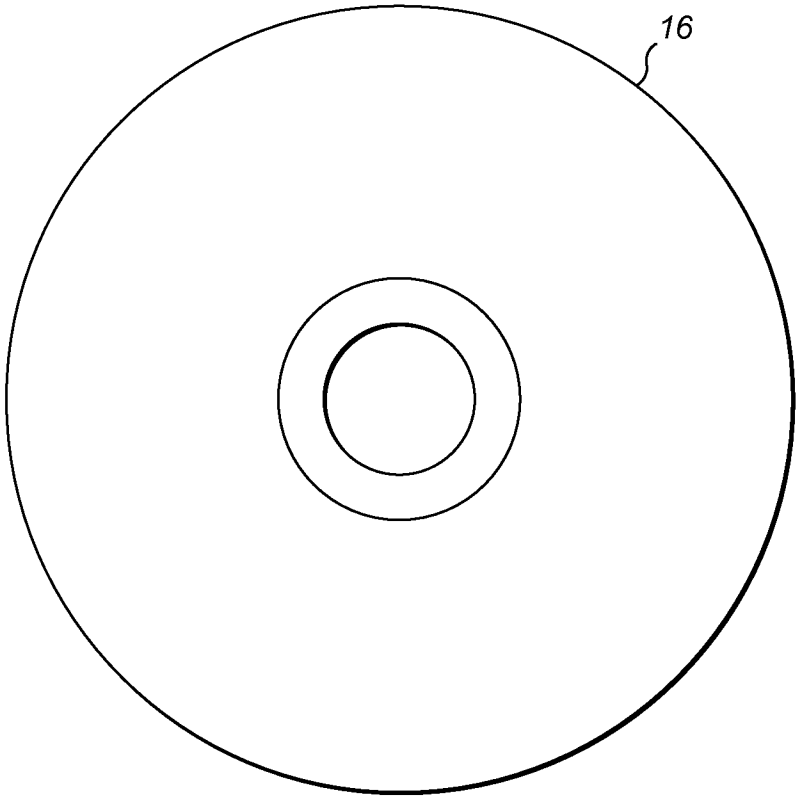

FIG. 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/FI2018/050001

A. **CLASSIFICATION OF SUBJECT MATTER**

See extra sheet

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)

IPC: G06F, G06K, H04N

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

FI, SE, NO, DK

Electronic data base consulted during the international search (name of data base, and, where practicable, search terms used)

EPODOC, WPIAP, EPO-Internal full-text databases, BIOSIS, COMPDX, EMBASE, INSPEC, MEDLINE, NPL, TDB, XP3GPP, XPAIP, XPESP, XPETSI, XPI3E, XPIEE, XPIETF, XPIOP, XPIPCOM, XPJEPG, XPMISC, XPOAC, XPRD, XPTK

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	YU, YOUNGJAE et al., 'End-to-end Concept Word Detection for Video Captioning, Retrieval, and Question Answering'. In: arXiv.org, Cornell University Library [online], 2016-12-13, arXiv:1610.02947v2, [retrieved on 2018-04-16]. Retrieved from <https://arxiv.org/pdf/1610.02947v2> abstract; section 1. 'Introduction'; section 1.1, subsection 'Attention for Captioning'; sections 1.2, 2.1 - 2.2, 3.1, 4.1; appendix section A.2; figure 11	1 – 14

☒ 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
19 April 2018 (19.04.2018)	23 April 2018 (23.04.2018)

Name and mailing address of the ISA/FI
Finnish Patent and Registration Office
FI-00091 PRH, FINLAND
Facsimile No. +358 29 509 5328

Authorized officer
Antti Salmela
Telephone No. +358 29 509 5000

INTERNATIONAL SEARCH REPORT

International application No.
PCT/FI2018/050001

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	TAPASWI, MAKARAND et al., 'MovieQA: Understanding Stories in Movies through Question-Answering'. In: Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), 26 June - 1 July 2016, Las Vegas, NV, USA. Los Alamitos, CA, USA: IEEE Computer Society, 2016-12-12, pp. 4631 - 4640, ISSN 1063-6919, ISBN 978-1-4673-8851-1, [retrieved on 2018-04-17], <DOI:10.1109/CVPR.2016.501>, XP 033021654 abstract; section 2, subsection 'Video understanding via language'; section 3. 'MovieQA dataset'; section 3.2, paragraph 1; section 4. 'Multi-choice Question-Answering'; sections 4.2 - 4.4, 5.4	1 – 14
X	SHARGHI, AIDEAN et al., 'Query-Focused Extractive Video Summarization'. In: Computer Vision - ECCV 2016, Lecture Notes in Computer Science, edited by LEIBE, B. et al. Cham, Switzerland: Springer, 2016-09-17, Vol. 9912, pp. 3 - 19, ISBN 978-3-319-46484-8, [retrieved on 2018-04-17], <DOI:10.1007/978-3-319-46484-8_1>, XP 047362379 abstract; section 1, paragraph 2; page 5, paragraph 3; sections 4.1 - 4.6, 5.2	1 – 14
X	HU, WEIMING et al., 'A Survey on Visual Content-Based Video Indexing and Retrieval'. In: IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). IEEE, 2011-03-10, Vol. 41, No. 6, pp. 797 - 819, ISSN 1094-6977, [retrieved on 2018-04-17], <DOI:10.1109/TSMCC.2011.2109710>, XP 011479468 abstract; section I, paragraphs 5 - 6; sections III.B; section III.C, item 2; section IV.C; section V.A, items 4 - 5; section V.B, items 2 - 4; section VI, paragraphs 1 - 3; section VI.C; section VII, items 5, 8 - 9; section VIII	1 – 14
X	TU, KEWEI et al., 'Joint Video and Text Parsing for Understanding Events and Answering Queries'. In: arXiv.org, Cornell University Library [online], 2014-02-21, arXiv:1308.6628v2, [retrieved on 2018-04-16]. Retrieved from <https://arxiv.org/pdf/1308.6628v2> abstract; sections 1.1 - 1.3, 3.3; section 4 'Text Parsing'; sections 5.3, 6, 6.1 - 6.3	1 – 14
X	US 2013282747 A1 (CHENG HUI [US] et al.) 24 October 2013 (24.10.2013) abstract; figures 1, 8; paragraphs [0031] - [0042], [0046], [0048], [0055], [0066] - [0068], [0071] - [0075]	1 – 14

INTERNATIONAL SEARCH REPORT

International application No.
PCT/FI2018/050001

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	TAPASWI, MAKARAND et al., 'Aligning plot synopses to videos for story-based retrieval'. In: International Journal of Multimedia Information Retrieval (2015). London: Springer, 2014-09-11, Vol. 4, No. 1, pp. 3 - 16, ISSN 2192-662X, [retrieved on 2018-04-16], <DOI:10.1007/s13735-014-0065-9> abstract; section 1, paragraphs 4 - 6; section 2.1, paragraph 2; sections 2.2 - 2.3, 5, 5.1, 6.4, 7	1 – 14
X	LIN, DAHUA et al., 'Visual Semantic Search: Retrieving Videos via Complex Textual Queries'. In: Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014), 23 - 28 June 2014, Columbus, OH, USA. Los Alamitos, CA, USA: IEEE Computer Society, 2014-09-25, pp. 2657 - 2664, ISSN 1063-6919, ISBN 978-1-4799-5118-5,], <DOI:10.1109/CVPR.2014.340>, XP 032649129 abstract; section 1, paragraphs 4 - 6; section 3; section 4. 'Visual Semantic Search'; sections 4.3, 5.1 - 5.2, 6	1 – 14
A	MUN, JONGHWAN et al., 'MarioQA: Answering Questions by Watching Gameplay Videos'. In: arXiv.org, Cornell University Library [online], 2016-12-06, arXiv:1612.01669v1, [retrieved on 2018-04-11]. Retrieved from <https://arxiv.org/pdf/1612.01669v1>, XP 080737096 abstract; section 1, paragraph 3; section 2, paragraph 2; sections 3.1, 4.1 - 4.4, 5.2	

INTERNATIONAL SEARCH REPORT
Information on Patent Family Members

International application No.
PCT/FI2018/050001

Patent document cited in search report	Publication date	Patent family members(s)	Publication date
US 2013282747 A1	24/10/2013	US 9244924 B2	26/01/2016
		US 2013198197 A1	01/08/2013
		US 9053194 B2	09/06/2015
		US 2015254231 A1	10/09/2015
		US 9563623 B2	07/02/2017
		US 2014328570 A1	06/11/2014
		US 2016004911 A1	07/01/2016
		US 2016110433 A1	21/04/2016
		US 2016154882 A1	02/06/2016
		US 2016328384 A1	10/11/2016

.....

INTERNATIONAL SEARCH REPORT

International application No.
PCT/FI2018/050001

CLASSIFICATION OF SUBJECT MATTER

IPC
G06F 17/30 (2006.01)
H04N 21/8549 (2011.01)
H04N 21/8405 (2011.01)
G06K 9/00 (2006.01)
G06N 3/04 (2006.01)