



- (51) International Patent Classification: *G06K 9/00* (2006.01) *G06F 3/16* (2006.01) LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR).
- (21) International Application Number: PCT/US2017/040358 **Published:**
— with international search report (Art. 21(3))
- (22) International Filing Date: 30 June 2017 (30.06.2017)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data: 62/361,160 12 July 2016 (12.07.2016) US
- (71) Applicant: **PROTEQ TECHNOLOGIES LLC** [US/US];
1000 N. West Street, Suite 1200, Wilmington, Delaware
19801 (US).
- (74) Agent: **BASILE, Andrew R., JR.** et al.; 3001 W. Big
Beaver Road, Suite 624, Troy, Michigan 48084 (US).
- (81) Designated States (unless otherwise indicated, for every
kind of national protection available): AT, AU, BG, BH,
BR, CA, CH, CN, CO, DE, EC, ES, FI, GB, HR, HU, ID,
IL, IN, IS, JP, KR, LU, LY, MA, MN, MX, MY, NA, NO,
NZ, PE, PG, PH, PL, QA, RO, RS, RW, SD, SE, SG, SM,
SV, SY, TN, US, VC, ZA.
- (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), European (AL, AT, BE, BG, CH, CY, CZ,
DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT,

(54) Title: INTELLIGENT SOFTWARE AGENT

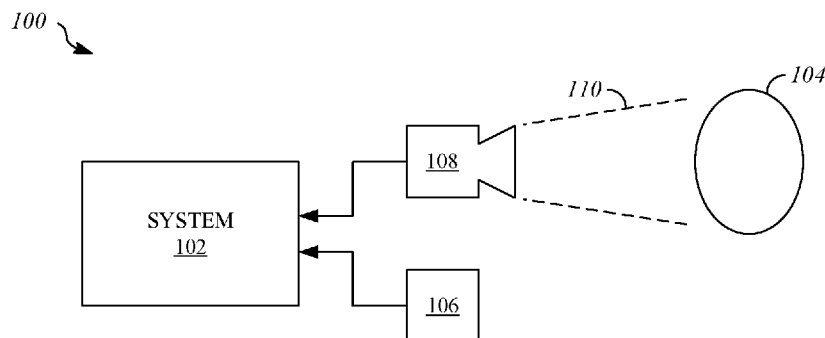


FIG. 1

(57) Abstract: Modifying operation of an intelligent agent in response to facial expressions and/or emotions.

WO 2018/013366 A1

INTELLIGENT SOFTWARE AGENT

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of United States Provisional Application No. 62/361,160 filed on July 12, 2016 and entitled “Intelligent Software Agent,” which is incorporated herein in its entirety by reference.

TECHNICAL FIELD

[0002] The application relates generally to intelligent software agents.

BACKGROUND

[0003] Intelligent software agents can perform actions on behalf of a user. Actions can be performed in response to a natural-language user input, such as a sentence spoken by the user. In some circumstances, an action taken by an intelligent software agent may not match the action that the user intended.

SUMMARY

[0004] One aspect of the disclosed embodiments is a system. The system includes a microphone configured to receive an audio input, a camera configured to obtain one or more images, and one or more processors. The one or more processors are configured to perform a first action in response to the audio input, identify information representing a facial expression in the one or more images, determine whether the identified information representing the facial expression corresponds to a first reaction or a second reaction to the performing of the first action in response to the audio input, and in accordance with a determination that the identified information representing the facial expression corresponds to the second reaction, perform a second action different from the first action.

[0005] Another aspect of the disclosed embodiments is a method. The method includes obtaining, by a microphone, an audio input, and obtaining, by a camera, one or more images. The method also includes performing, by one or more processors, a first action in response to the audio input, and identifying, by the one or more processors, information representing a facial expression in the one or more images. The method also includes determining, by the one or more processors, whether the identified information representing the facial expression corresponds to a first reaction or a second reaction to the performing of the first action in response to the command input, and in accordance with a determination that the identified

information representing the facial expression corresponds to the second reaction, performing a second action different from the first action.

[0006] Another aspect of the disclosed embodiments is a non-transitory computer-readable storage device including program instructions executable by one or more processors. The program instructions, when executed, cause the one or more processors to perform operations. The operations include obtaining, from a microphone, an audio input, and obtaining, from a camera, one or more images. The operations also include performing a first action in response to the audio input, and identifying information representing a facial expression in the one or more images. The operations also include determining whether the identified information representing the facial expression corresponds to a first reaction or a second reaction to the performing of the first action in response to the audio input, in accordance with a determination that the identified information representing the facial expression corresponds to the second reaction, performing a second action different from the first action, and in accordance with a determination that the identified information representing the facial expression corresponds to the first reaction, forgo performing the second action.

BRIEF DESCRIPTION OF THE DRAWINGS

- [0007]** FIG. 1 is an illustration showing an exemplary environment.
- [0008]** FIG. 2 is an illustration showing an exemplary system.
- [0009]** FIG. 3 is an illustration showing analysis and classification of expression signals.
- [0010]** FIG. 4 is a flowchart showing an exemplary process.
- [0011]** FIG. 5 is a flowchart showing an exemplary process.
- [0012]** FIG. 6 is a flowchart showing an exemplary process.
- [0013]** FIG. 7 is a flowchart showing an exemplary process.
- [0014]** FIG. 8 is an illustration showing an exemplary hardware configuration for a controller.

DETAILED DESCRIPTION

[0015] FIG. 1 shows an environment 100 in which an exemplary system 102 having an intelligent software agent may be utilized. The system 102 is operable to receive a command input from a user 104. The command input may be expressed in natural language, such as a sentence that expresses an intention. The system 102 interprets the command input, and determines an inferred intention. The inferred intention may represent an action that the

system 102 is able to take on the user's behalf. Thus, the system 102 may perform the action represented by the inferred intention.

[0016] In some embodiments, the command input is a speech input. The speech input may be signals and/or data that represent an utterance such as a spoken word, phrase, or sentence. The system 102 may receive an audio input signal that includes the speech input from an audio input device such as a microphone 106. In some embodiments, the command input is a text input. The system 102 may receive the text input from an input device such a physical keyboard or an on-screen keyboard. In some embodiments, the command input is expressed in a non-spoken language. As an example, the command input may be expressed in a sign language, such as American Sign Language. The non-spoken command input may be received by the system 102 using a visual input device that is operable to output one or more images, such as a video camera 108. A field of view 110 of the video camera 108 may be oriented to capture images representing a portion of the body of the user 104, such as the user's face and/or the user's hands.

[0017] FIG. 2 shows components of the system 102. In the illustrated embodiment, a command input 220 is received as an input at a command analyzer 222. The command analyzer 222 may be a software module that is operable to analyze the command input 220, identify one or more portions of the command input 220 that can be converted into text or tokens, such as spoken or non-spoken language, and convert the command input into a form that can be understood by other components of the system 102, such as text or tokens. In some embodiments, the command analyzer 222 is a text-to speech convertor that converts a speech input into text. Many suitable methods are known for converting speech inputs into text, such as Hidden Markov Models and deep feedforward neural networks. In some embodiments, the speech analyzer provides the speech input converted into words or tokens as an output. In some embodiments, the command analyzer 222 utilizes machine-vision techniques to perceive non-spoken communication in an image or a series of images such as video frames, and provides the non-spoken command input converted into words or tokens as an output. In implementations where the command input 220 is a text input, no action may be needed by the command analyzer 222.

[0018] The system 102 includes an intention analyzer 224 that receives the command input subsequent to conversion by the command analyzer 222. The command input 220 may be received by the intention analyzer 224 as a text-based input. The intention analyzer 224 is operable to determine a user intention based on the command input and to select an action to be performed in response to the command input.

[0019] In some embodiments, the intention analyzer 224 identifies a relationship between the command input 220 and an action that can be performed by the system 102.

Relationships between the command input 220 and actions can be identified using previously stored information. In some embodiments, the system 102 includes a storage device 226 where this and other items of information may be stored. In some embodiments, the intention analyzer 224 determines the user intention and/or selects the action by identifying a semantic relationship model that defines a relationship between the command input 220 and the action.

[0020] In a simple example, the intention analyzer 224 identifies the relationship between the command input and the action based on the presence of one or more words or tokens in the command input. In some embodiments, at least some of the words or tokens are classified into categories, and these categories are used to identify the relationship between the words or tokens and the action. In some embodiments, relationships between the command input and the actions are identified using an ontology that represents actions that can be performed by the system 102 and relates those actions to properties that may be present in the command input 220.

[0021] In some embodiments, the intention analyzer 224 attempts to identify user intentions that are related to actions that can be performed by the system 102. These user intentions are typically represented in the command input by a word or by a series of words. Thus, the intention analyzer 224 may identify actions based on a relationship between the command input 220 and the actions. In some embodiments, this identification is performed using previously stored information that defines semantic relationships between command inputs or portions of command inputs and actions, such as the ontology.

[0022] Each semantic relationship may have a strength that describes how weak or strong the relationship between the command input 220 and the user intention or action is. As an example, the strength of each semantic relationship may be expressed as a probability that a particular command input corresponds to a particular user intention or action to be performed. Semantic relationships may be strengthened or weakened by information received by the system, such as feedback information received after performing the action, as will be described herein.

[0023] The system 102 may include an action processor 228. The action processor 228 is operable to perform actions, such as by performing a first action in response to the command input. In some embodiments, the action processor receives information from the intention analyzer 224 that identifies an action to be performed. In response to receiving this information, the action processor 228 determines how to cause the desired action to be

performed. The action processor 228 may cause the desired action may be performed by the system 102. As an example, the system 102 may be able to answer a question using information previously stored at the storage device 226. The action processor 228 may cause the desired action to be performed by an external service 230 that is accessed, for example, using an application programming interface (API). As examples, the external services 230 may include messaging, telephony, and calendar services.

[0024] The output unit 232 is operable to create a message for communication to the user and to communicate the message as output 234. As an example, the output 234 may be a message that responds to the command input by announcing, in the form of text or audible synthesized speech, the action that will be performed by the system 102 in response to the command input 220.

[0025] To receive feedback from the user without prompting the user, the system 102 may include an expression analyzer 242 and a classifier 244. A video input 240 is received as an input by the expression analyzer 242. The video input 240 may be obtained, for example from a video input device such as the video camera 108. The video input 240 may be, as examples, a portion of an image, an image, or a sequence of images.

[0026] The expression analyzer 242 may use facial recognition techniques to identify a human face in the video input 240, and to determine the identity of the user based on information that is, for example, accessed from the storage device 226. By determining the identity of the user, the expression analyzer 242 and/or other portions of the system 102 may access information associated with the user, such as a user profile.

[0027] The expression analyzer 242 is operable to recognize shapes and/or movement patterns in the video input 240 that correspond to portions of facial expressions. As an example, the face image in the video input 240 may be analyzed to determine whether particular muscles or muscle groups are activated by identifying shapes or motions that are associated with activation of a particular muscle or muscle group. Each of these shapes or movement patterns may be used by the expression analyzer to set or modify the strength of a signal that corresponds to a particular type or aspect of facial expression, and may be referred to herein as a facial expression signal. Since aspects of the face image in the video input 240 may correspond to multiple types of facial expressions, the expression analyzer 242 may output multiple facial expression signals each having a signal strength corresponding to a degree to which a particular type of facial expression is present in the video input 240. In some embodiments, the expression analyzer 242 utilizes automated emotion recognition techniques. In some embodiments, automated emotion recognition techniques may be

implemented using a taxonomy of human facial movements categorized by their appearance on the face, such as the Facial Action Coding System.

[0028] The classification may represent a current reaction or emotional state of the user. The classification may be determined by comparing the facial expression signals output by the classifier 244 with information that is based on previously observed facial expression signals and known states. In some embodiments, machine-learning techniques are utilized to classify the facial expression signals. As examples, the classifier 244 may be implemented using a neural network, a deep neural network, or a support vector machine.

[0029] The classifier 244 may be operable to determine a current emotional state for the user. In some embodiments, the current emotional state determined by the classifier 244 is associated with a metric that describes a probability that the emotional state determined by the classifier 244 correctly represents the user's current emotional state. In other embodiments, the current emotional state determined by the classifier 244 is associated with a metric that describes a degree to which the user is experiencing the emotional state determined by the classifier 244.

[0030] In some embodiments, the system 102 determines a user reaction to the action performed by the action processor 228 or to the output 234. The user reaction may be represented by user reaction information. In some embodiments, the user reaction information is used to classify the reaction by category, such as a positive or negative reaction. In other embodiments, the user reaction information is expressed as one or more metrics such as a probability that the user reaction corresponds to a certain state such as positive or negative, or a degree to which the user is expressing the reaction. As used herein, determining that a reaction is positive may include determining that a probability of a positive reaction is above a threshold, and determining that a reaction is negative may include determining that a probability of a negative reaction is above a threshold.

[0031] The user reaction may be determined by the classifier 244 in the manner described with respect to determining the user's emotional state. However, the user reaction is measured during a time period that follows the action or output 234 of the system 102. In some embodiments, the start of the time periods in which the system 102 starts measuring the user reaction corresponds to the time at which the system 102 performs the action or the output 234. Thus, the action or output 234 is used as a trigger that starts monitoring and/or measuring of the user reaction.

[0032] The duration of the time period is chosen to represent the time in which the user may be expressing a reaction to the action or the output 234. In some embodiments, the time

period is a predetermined length of time. In some embodiments, the time period is determined dynamically on a per-user basis based on previously observed reactions by the user. Accordingly, monitoring and/or measuring of the user reaction may end a predetermined time after performance of the action or the output 234 by the system 102.

[0033] In some embodiments, the classifier 244 may also be operable to analyze and classify emotional responses present in the command input 220, whether in the form of audio, video, or text, and the user reaction determined from the command input 220 may be utilized in the same manner described with respect to the video input 240.

[0034] As shown in FIG. 3, aspects of the appearance of the user's face may be categorized according to a plurality of expression types 346 by the expression analyzer 242. The expression types 346 may be classifications that describe positions and/or movements of the user's face and/or emotional states indicated by the positions and/or movements of the user's face. In the illustrated example, the expression types 346 are represented as EXP_1 through EXP_n. For each expression type, the expression analyzer 242 may output a respective facial expression signal 348, which changes over time in accordance with movements of the user's face.

[0035] The classifier 244 receives the facial expression signals 348 from the expression analyzer 242, as represented by μ_1 through μ_n in FIG. 3. Using the facial expression signals 348, the classifier determines a classification 352 for the user's current facial expression.

[0036] Normalizing 350 may optionally be applied to the facial expression signals 348 prior to determining the classification. Normalizing 350 may include applying a non-linear normalization function. In some embodiments, the facial expression signals 348 may be normalized relative to a baseline. In some embodiments, the baseline may be a global baseline that is based on information obtained from other users' facial expressions. In some embodiments, the baseline may be user-specific, based in part or in whole on data obtained from the user. In some embodiments, normalizing 350 may be performed by applying gain values to the facial expression signals 348 from the expression analyzer 244, as represented by values gain_1 through gain_n in FIG. 3. In some implementations, the gain values are learned by the classifier 244 using machine learning techniques. In some implementations, normalizing 350 also includes applying an offset value to each of the facial expression signals.

[0037] The classification 352 is determined by the classifier 244 using the facial expression signals 348 subsequent to normalizing 350. The classification may be determined by the classifier 244 by using the normalized facial expression signals as inputs to a comparison or a model that is based on previously observed facial expression signals and known states, as previously described.

[0038] It should be understood that the functions and components of the system 102 may be implemented using multiple computing devices. In one embodiment, a first device implements a front end user interface, receives inputs, and generates outputs, while operations such as converting speech to text or selecting an action are performed using a second device, which may be a remote device that is accessed using a network.

[0039] FIG. 4 shows an exemplary process 400. The process 400 may be performed, for example, by the system 102. Operation 410 includes obtaining a command input from a user. The command input may be obtained using an input device such as a keyboard, a microphone, or a camera. In some embodiments, the command input may be a speech input that is obtained from an audio input device such as the microphone 106. In some embodiments, the command input is obtained by receiving information that represents the command input, such as by a transmission over a network or by accessing the information from a storage device.

[0040] In some embodiments, obtaining the command input in operation 410 further comprises operation 412, in which the command input is converted into text, such as by converting speech that is present in the audio input into text. Converting the command input into text in operation 412 may be performed, for example, as described with respect to the command analyzer 222.

[0041] Operation 420 includes selecting a first action based on the command input. Operation 420 may be performed, for example, in the manner described with respect to the intention analyzer 224.

[0042] In some embodiments, operation 420 further comprises identifying a user intention in operation 422 based on the command input. Identifying a user intention may be performed, for example, in the manner described with respect to the intention analyzer 224.

[0043] In some embodiments, operation 420 further comprises operation 424 and operation 426. In operation 424, one or more semantic relationship models are identified. The one or more semantic relationship models define a relationship between the command input and one or more actions. As an example, numerous semantic relationship models can be identified that describe a relationship between the command input and actions. Each of these semantic

relationship models may be associated with a strength, which defines a probability that the command input represents a user intention to have the system 102 perform a particular action. In operation 426, a first action is selected using the semantic relationship models. As an example, the strongest (e.g., highest probability) semantic relationship model can be chosen for use by the system, resulting in selection of the action associated with that semantic relationship model being utilized as the first action.

[0044] In operation 430, the first action is performed in response to the command input. Performing the first action may include automated performance of one or more tasks on behalf of the user. The first action may be performed by the system 102 or by the external service 230 as directed by the system 102. Operation 430 can be performed, for example, as described with respect to the action processor 228.

[0045] FIG. 5 shows an exemplary process 500. Process 500 may be performed by the system 102 subsequent to process 400, for example, in order to obtain feedback resulting from performance of the first action in operation 430. In some embodiments, performance of the action in operation 430 is utilized as a start time for obtaining feedback from the user, since the system 102 is aware of the time at which the action is performed in operation 430.

[0046] Operation 510 includes obtaining a video input. As an example, the video input 240 may be received from a video input device such as the video camera 108.

[0047] Operation 520 includes identifying information representing a facial expression of the user in the video input. Identifying information representing a facial expression may be performed as described with respect to the expression analyzer 242.

[0048] In some embodiments, operation 520 further comprises operation 522. In operation 522, the information representing the facial expression of the user in video input includes a plurality of facial expression signals that each represent an aspect of the user's facial expression, and operation 522 includes normalizing the facial expression signals.

[0049] Operation 530 includes determining a user reaction. In some embodiments, determining a user reaction can be performed by comparing the information representing the facial expression of the user with stored information that describes relationships between facial expressions and reactions. Operation 530 is performed using the information obtained at operation 520 and optionally using the information obtained at operation 522. Operation 530 may utilize information that begins at a first time point when an action is performed by the system 102, such as the action performed at operation 430. The information may include a time period that extends over a predetermined time period that is static in duration or is determined dynamically based on user behavior.

[0050] In some embodiments, operation 530 further comprises operation 532. Operation 532 includes classifying the information representing the facial expressions of the user. Classifying the facial expressions of the user can be performed, for example, as described with respect to the classifier 244.

[0051] Operation 540 includes storing user reaction information that describes the user reaction. As an example, the system 102 may store the user reaction information using the storage device 226.

[0052] In some embodiments, process 500 includes operation 550. Operation 550 includes taking action in response to the user reaction information. As an example, operation 550 may include modifying an aspect of operation of the system 102 in response to the user's facial expressions. In some embodiments, modification of the system 102 performed in operation 550 may change the results generated by the process 400, such as by modifying the information used by the intention analyzer 224 to select the action based on the command input. This may cause the same user input to cause performance of a different action by the system 102 during subsequent usage of the system 102 by the user, as a result of the feedback information provided by the process 500.

[0053] As shown in the foregoing operations, by implementing the process 500, the system 102 may start monitoring a facial expression that is related to an action after performing the action, and continue monitoring the facial expression that is related to the action after a predetermined time period. The system 102 may then store and/or action on the information obtained from the facial expression.

[0054] FIG. 6 shows an exemplary process 600. Process 600 may be performed, for example, as part of or in lieu of taking action in response to the user reaction information in operation 550 of process 500.

[0055] Operation 610 includes determining that the user reaction information indicates a negative reaction. As explained with respect to the classifier 244, this can be performed by, for example, classifying user reaction information based on known user reaction states that were previously made available to the system 102, and may be expressed as a category or a probability that the user's reaction is a negative reaction.

[0056] Operation 620 includes identifying a second action based on the command input and the user reaction information. The second action is identified in the same manner that the first action was identified. However, the system 102 selects the second action with knowledge that the first action was an incorrect response to the command input. In some embodiments, the system 102 chooses an action that, at the time of selection of the first

action, was ranked as being the second most likely user-intended action. In other embodiments, the second action is selected in the manner described with respect to operation 426, but subsequent to modification of information used by the system 102 to select actions, such as the semantic relationship models.

[0057] In some embodiments, the second action is identified in operation 620 based on a second command input that is received from the user immediately after the system 102 performs the first action. The second command input may repeat a request that was present in the first command input, either in identical phrasing or in modified phrasing. The request is an expression of an intention for the system 102 to respond in a particular way. The request may be expressed as a verbal command. There may be many different verbal commands that can be formed to express the same intention, and these various expressions are considered the same request. Thus, in some embodiments, a first command input represents a request, and a second command input received from the user after the system performs the first action represents the same request. The system 102 can determine that repetition of the request indicates that the first action did not correspond to the user's intention, and therefore is a negative reaction. In response, to the second command input and the determination that repetition of the request corresponds to a negative reaction, the system 102 can perform the second action in response to the second command input.

[0058] In operation 630, the second action is performed by the system. To collect further feedback information, the process 500 may be performed subsequent to performing the second action to determine whether the second action was correct or incorrect.

[0059] In some embodiments, if the system 102 determines, at operation 610, that the user's reaction is a positive reaction, the system 102 may, in accordance with the determination that the user reaction corresponds to a positive reaction, forgo performance of the second action by omitting operations 620 and 630.

[0060] FIG. 7 shows an exemplary process 700. Process 700 may be performed, for example, as part of or in lieu of taking action in response to the user reaction information in operation 550 of process 500.

[0061] Operation 710 includes modifying a semantic relationship between the command input and the action based on the user reaction information.

[0062] In some embodiments, operation 710 includes operation 712 and operation 714. Operation 712 includes determining that the user reaction information indicates a negative reaction and is similar to operation 620. In response to determining that the user reaction

information indicates the negative reaction, operation 714 includes weakening the semantic relationship model between the command input and the action.

[0063] In some embodiments, operation 710 includes operation 716 and operation 718. Operation 716 includes determining that the user reaction information indicates a positive reaction and is similar to operation 712 except that operation 712 tests for the opposite reaction. In response to determining that the user reaction information indicates the positive reaction, operation 718 includes strengthening the semantic relationship model between the command input and the action.

[0064] FIG. 8 shows an exemplary hardware configuration for a controller 800 that may be used to implement the system 102. The controller 800 may include a processor 810, memory 820, a storage device 830, one or more input devices 840, and one or more output devices 850. The controller 800 may include a bus 860 or a similar device to interconnect the components for communication. The processor 810 is operable to execute computer program instructions and perform operations described by the computer program instructions. As an example, the processor 810 may be a conventional device such as a central processing unit. The memory 820 may be a volatile, high-speed, short-term information storage device such as a random-access memory module. The storage device 830 may be a non-volatile information storage device such as a hard drive or a solid-state drive. The input devices 840 may include any type of human-machine interface such as buttons, switches, a keyboard, a mouse, a touchscreen input device, a gestural input device, or an audio input device. The output devices 850 may include any type of device operable to provide an indication to a user regarding an operating state, such as a display screen or an audio output.

CLAIMS

What is claimed is:

1. A system comprising:
a microphone configured to receive an audio input;
a camera configured to obtain one or more images; and
one or more processors configured to:
perform a first action in response to the audio input,
identify information representing a facial expression in the one or more images,
determine whether the identified information representing the facial expression corresponds to a first reaction or a second reaction to the performing of the first action in response to the audio input, and
in accordance with a determination that the identified information representing the facial expression corresponds to the second reaction, perform a second action different from the first action.
2. The system of claim 1, wherein the one or more processors are configured to:
in accordance with a determination that the identified information corresponds to the first reaction, forgo performance of the second action.
3. The system of claim 2, wherein the audio input represents a request, wherein the microphone is configured to receive a second audio input, the second audio input comprises the request, and wherein the one or more processors is configured to perform the second action in response to the second audio input and the determination that the identified information representing the facial expression corresponds to the second reaction.
4. The system of claims 1, wherein the one or more processors are further configured to:
analyze the audio input by obtaining a semantic relationship model that describes a semantic relationship of the audio input and the first action.
5. The system of claim 4, wherein the one or more processors are further configured to:

update the semantic relationship model using the identified information representing the facial expression.

6. The system of claim 5, wherein the one or more processors are further configured to: update the semantic relationship model to weaken the semantic relationship of the audio input and the first action in accordance with a determination that the identified information representing the facial expression corresponds to the second reaction.

7. The system of claim 5, wherein the one or more processors are further configured to: update the semantic relationship model to strengthen the semantic relationship of the audio input and the first action in accordance with a determination that the identified information representing the facial expression corresponds to the first reaction.

8. A method, comprising:
obtaining, by a microphone, an audio input;
obtaining, by a camera, one or more images;
performing, by one or more processors, a first action in response to the audio input;
identifying, by the one or more processors, information representing a facial expression in the one or more images;
determining, by the one or more processors, whether the identified information representing the facial expression corresponds to a first reaction or a second reaction to the performing of the first action in response to the audio input; and
in accordance with a determination that the identified information representing the facial expression corresponds to the second reaction, performing a second action different from the first action.

9. The method of claim 8, further comprising:
in accordance with a determination that the identified information representing the facial expression corresponds to the first reaction, forgo performing the second action.

10. The method of claim 8, wherein the audio input represents a request, the method further comprising:
obtaining, by the microphone, a second audio input, wherein the second audio input comprises the request; and

performing the second action in response to the second audio input and the determination that the identified information representing the facial expression corresponds to the second reaction.

11. The method of claim 8, further comprising:

selecting, by the one or more processors, the first action by identifying a semantic relationship model that describes a semantic relationship of the audio input and the first action.

12. The method of claim 11, further comprising:

updating, by the one or more processors, the semantic relationship model using the identified information representing the facial expression.

13. The method of claim 12, wherein updating the semantic relationship model using the identified information representing the facial expression comprises:

in accordance with a determination that the identified information representing the facial expression corresponds to the second reaction, weakening the semantic relationship of the audio input and the first action.

14. The method of claim 12, wherein updating the semantic relationship model using the identified information representing the facial expression comprises:

in accordance with a determination that the identified information representing the facial expression corresponds to the first reaction, strengthening the semantic relationship of the audio input and the first action.

15. A non-transitory computer-readable storage device including program instructions executable by one or more processors that, when executed, cause the one or more processors to perform operations, the operations comprising:

obtaining, from a microphone, an audio input;

obtaining, from a camera, one or more images;

performing a first action in response to the audio input;

identifying information representing a facial expression in the one or more images;

determining whether the identified information representing the facial expression corresponds to a first reaction or a second reaction to the performing of the first action in response to the audio input;

in accordance with a determination that the identified information representing the facial expression corresponds to the second reaction, performing a second action different from the first action; and

in accordance with a determination that the identified information representing the facial expression corresponds to the first reaction, forgo performing the second action.

16. The non-transitory computer-readable storage device of claim 15, wherein the audio input represents a request, the operations further comprising:

obtaining, by the microphone, a second audio input, wherein the second audio input comprises the request; and

performing the second action in response to the second audio input and the determination that the identified information representing the facial expression corresponds to the second reaction.

17. The non-transitory computer-readable storage device of claim 15, the operations further comprising:

selecting, by the one or more processors, the first action by identifying a semantic relationship model that describes a semantic relationship of the audio input and the first action.

18. The non-transitory computer-readable storage device of claim 15, the operations further comprising:

updating, by the one or more processors, the semantic relationship model using the identified information representing the facial expression.

19. The non-transitory computer-readable storage device of claim 18, wherein updating the semantic relationship model using the identified information representing the facial expression comprises:

in accordance with a determination that the identified information representing the facial expression corresponds to the second reaction, weakening the semantic relationship of the audio input and the first action.

20. The non-transitory computer-readable storage device of claim 18, wherein updating the semantic relationship model using the identified information representing the facial expression comprises:

in accordance with a determination that the identified information representing the facial expression corresponds to the first reaction, strengthening the semantic relationship of the audio input and the first action.

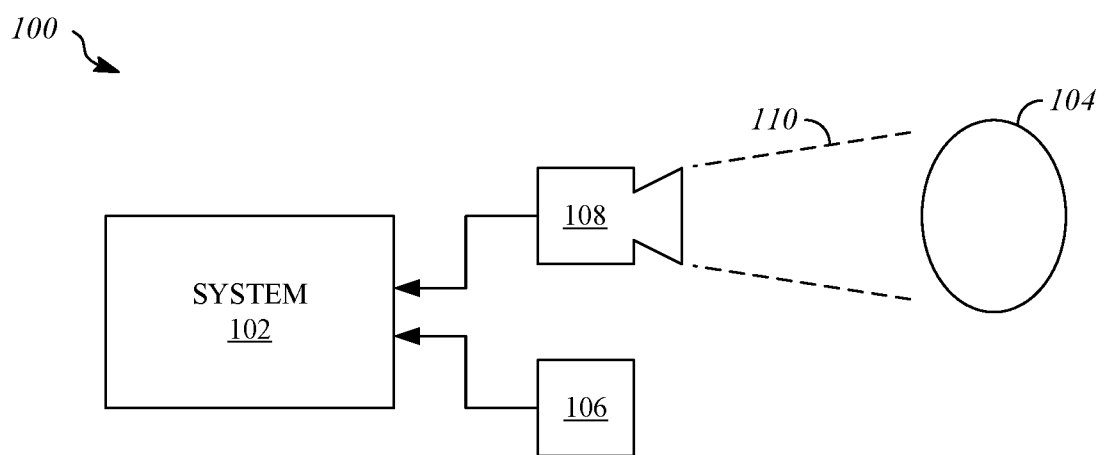


FIG. 1

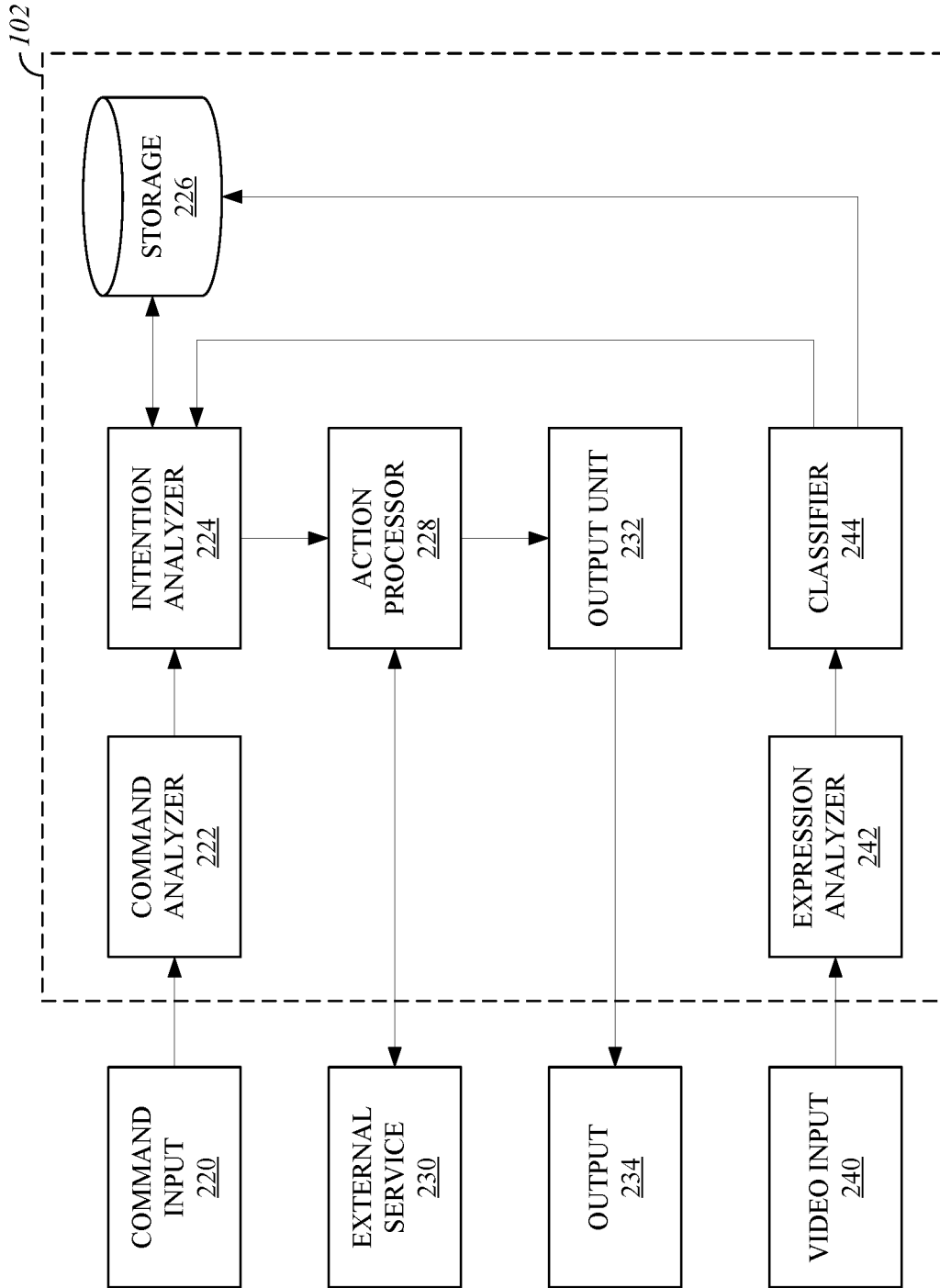


FIG. 2

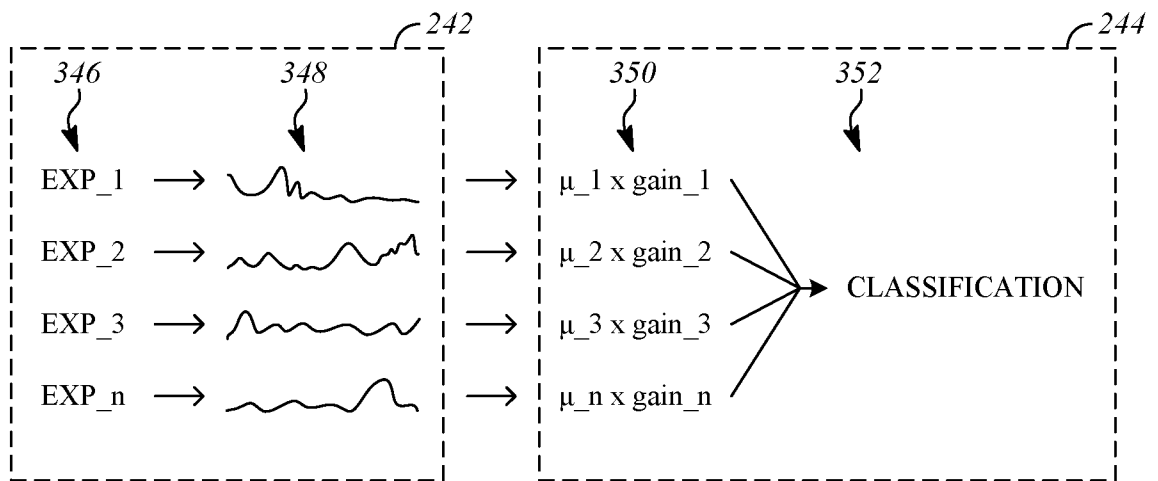


FIG. 3

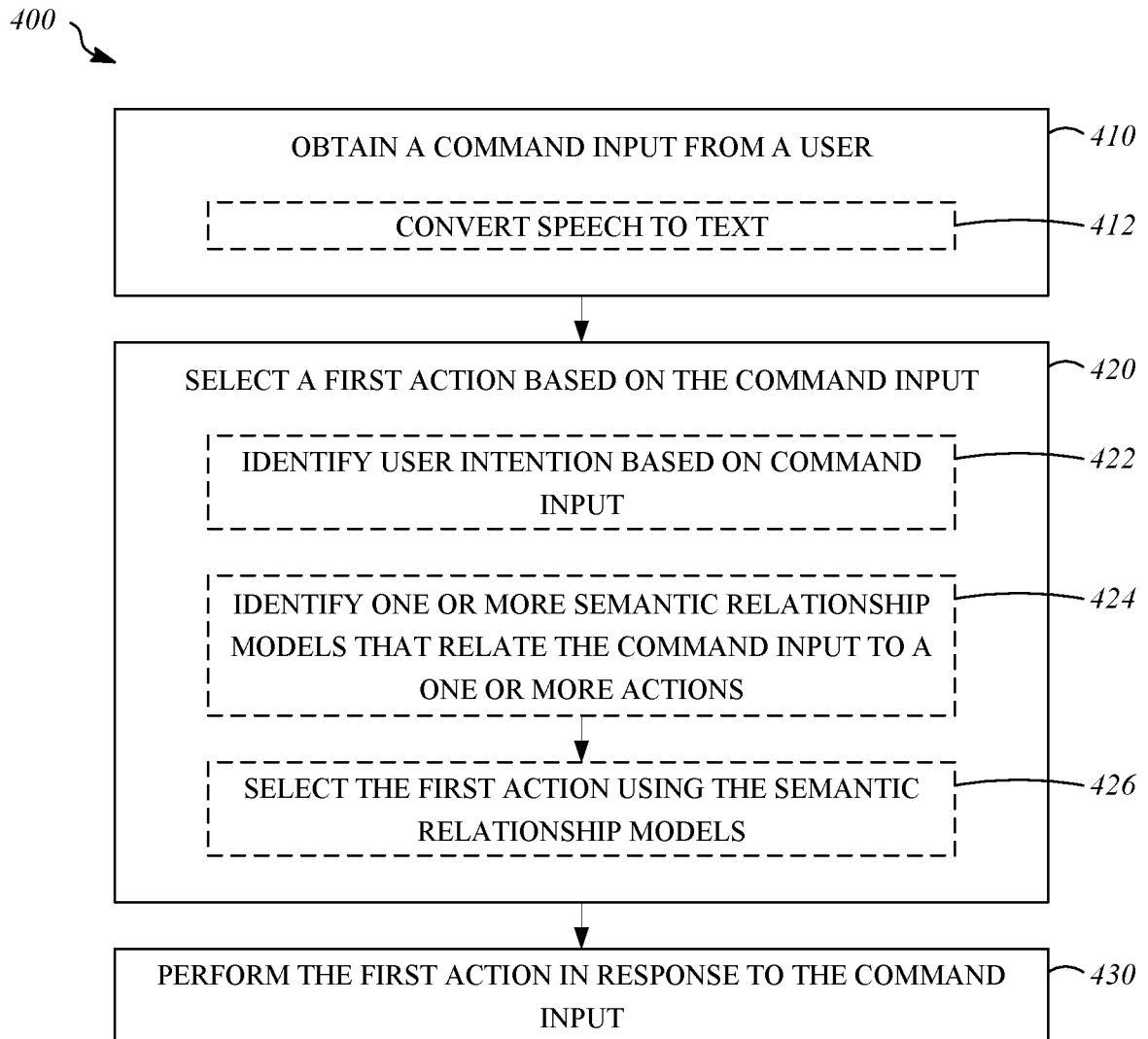


FIG. 4

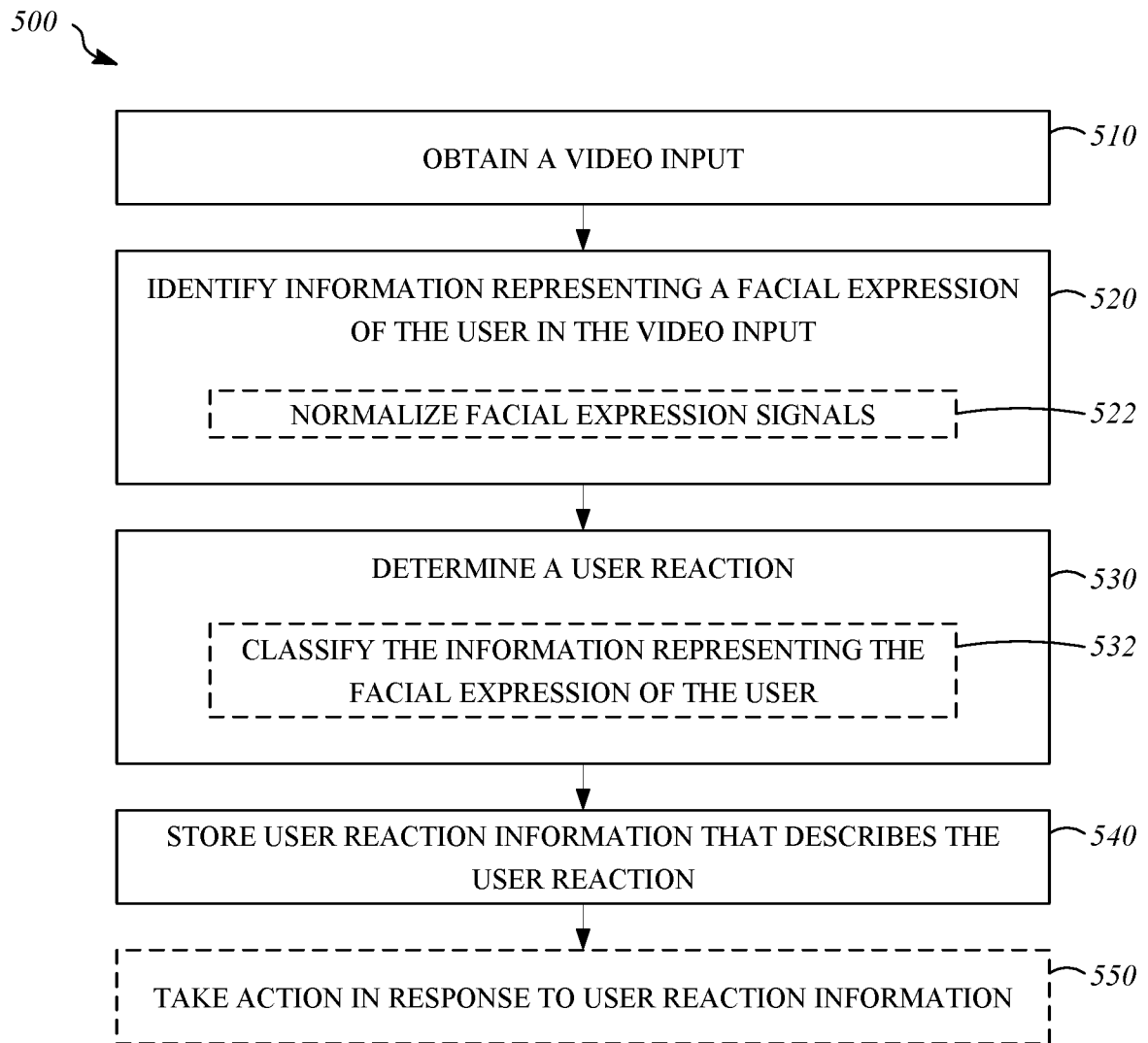
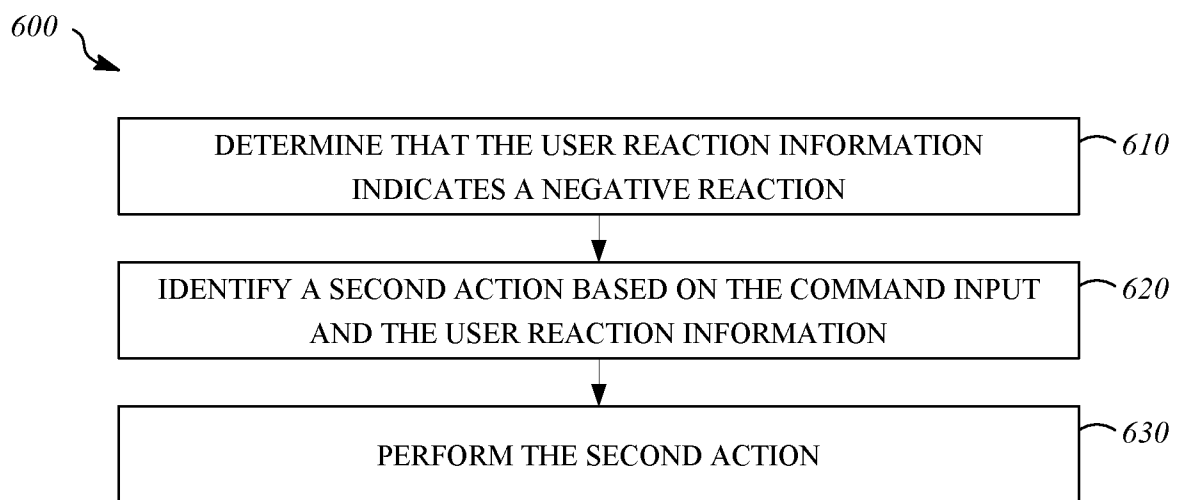


FIG. 5

**FIG. 6**

700 ↘

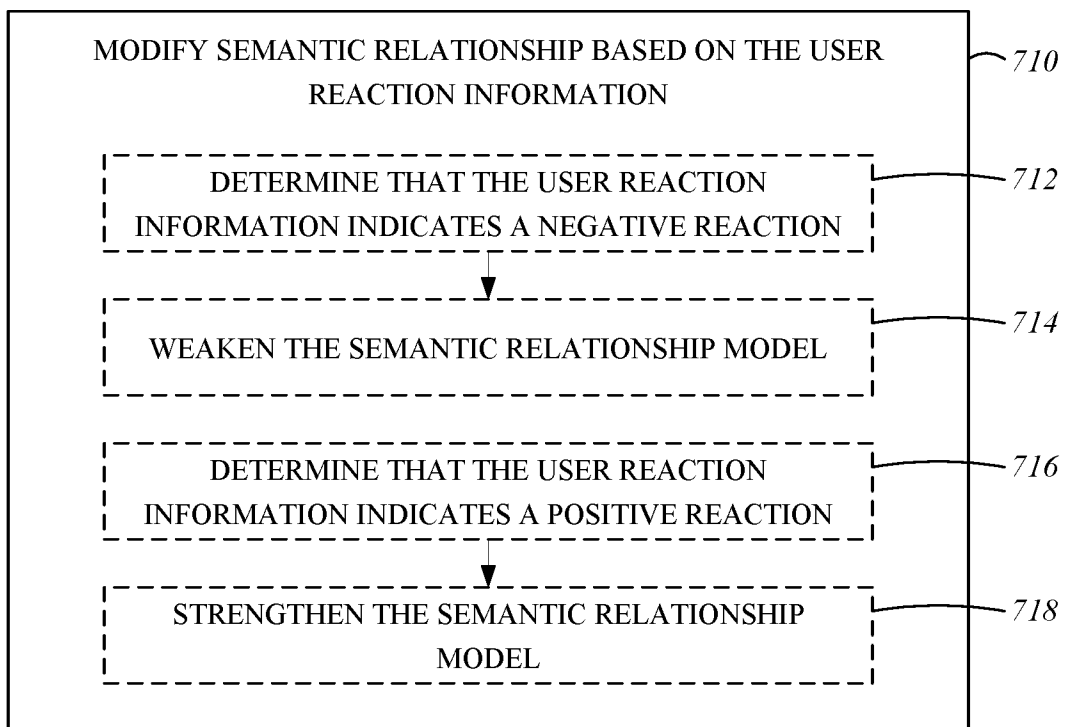


FIG. 7

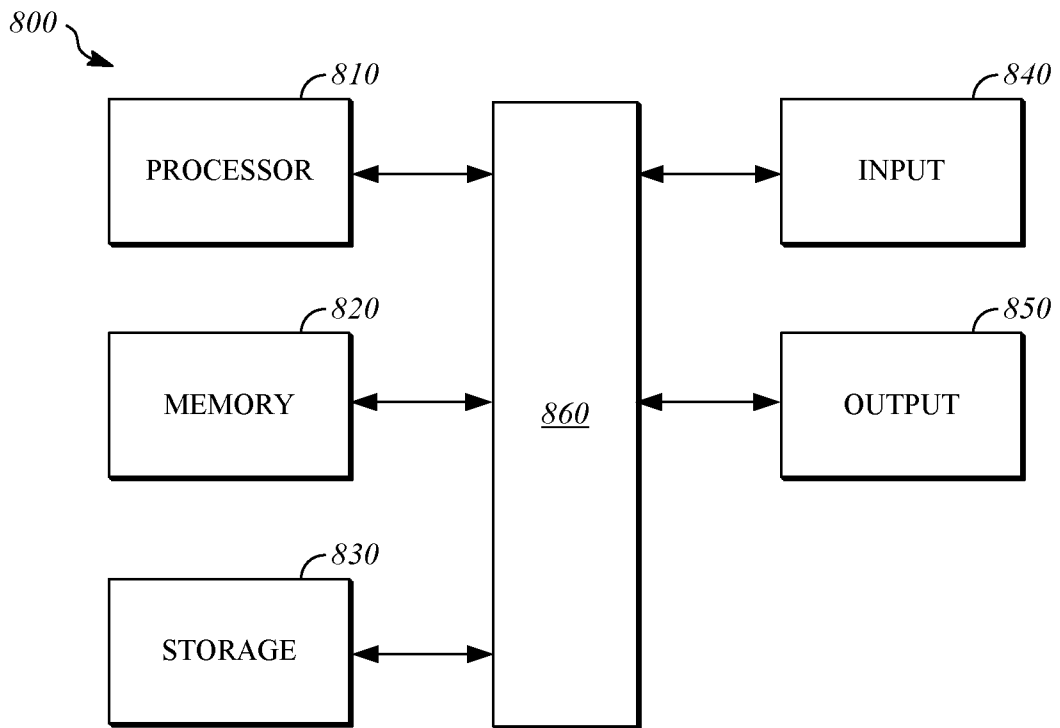


FIG. 8

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2017/040358

A. CLASSIFICATION OF SUBJECT MATTER
 INV. G06K9/00 G06F3/16
 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)
 G06K G06F
 Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2015/340031 A1 (KIM JUHEE [KR] ET AL) 26 November 2015 (2015-11-26) the whole document	1-20
A	US 2015/015690 A1 (ROH HYEONGSEOK [KR] ET AL) 15 January 2015 (2015-01-15) abstract paragraph [0132] - paragraph [0153]; figures 1,5	1-20
A	US 2014/139424 A1 (CHANG CHIH-SUNG [TW]) 22 May 2014 (2014-05-22) the whole document	1-20

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

27 September 2017

Date of mailing of the international search report

06/10/2017

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
 NL - 2280 HV Rijswijk
 Tel. (+31-70) 340-2040,
 Fax: (+31-70) 340-3016

Authorized officer

Coenen, Jean Pierre

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2017/040358

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2015340031 A1	26-11-2015	US 2015340031 A1 WO 2014109421 A1	26-11-2015 17-07-2014

US 2015015690 A1	15-01-2015	CN 104284013 A EP 2824538 A1 KR 20150007159 A US 2015015690 A1	14-01-2015 14-01-2015 20-01-2015 15-01-2015

US 2014139424 A1	22-05-2014	CN 103838366 A TW 201421376 A US 2014139424 A1	04-06-2014 01-06-2014 22-05-2014
