



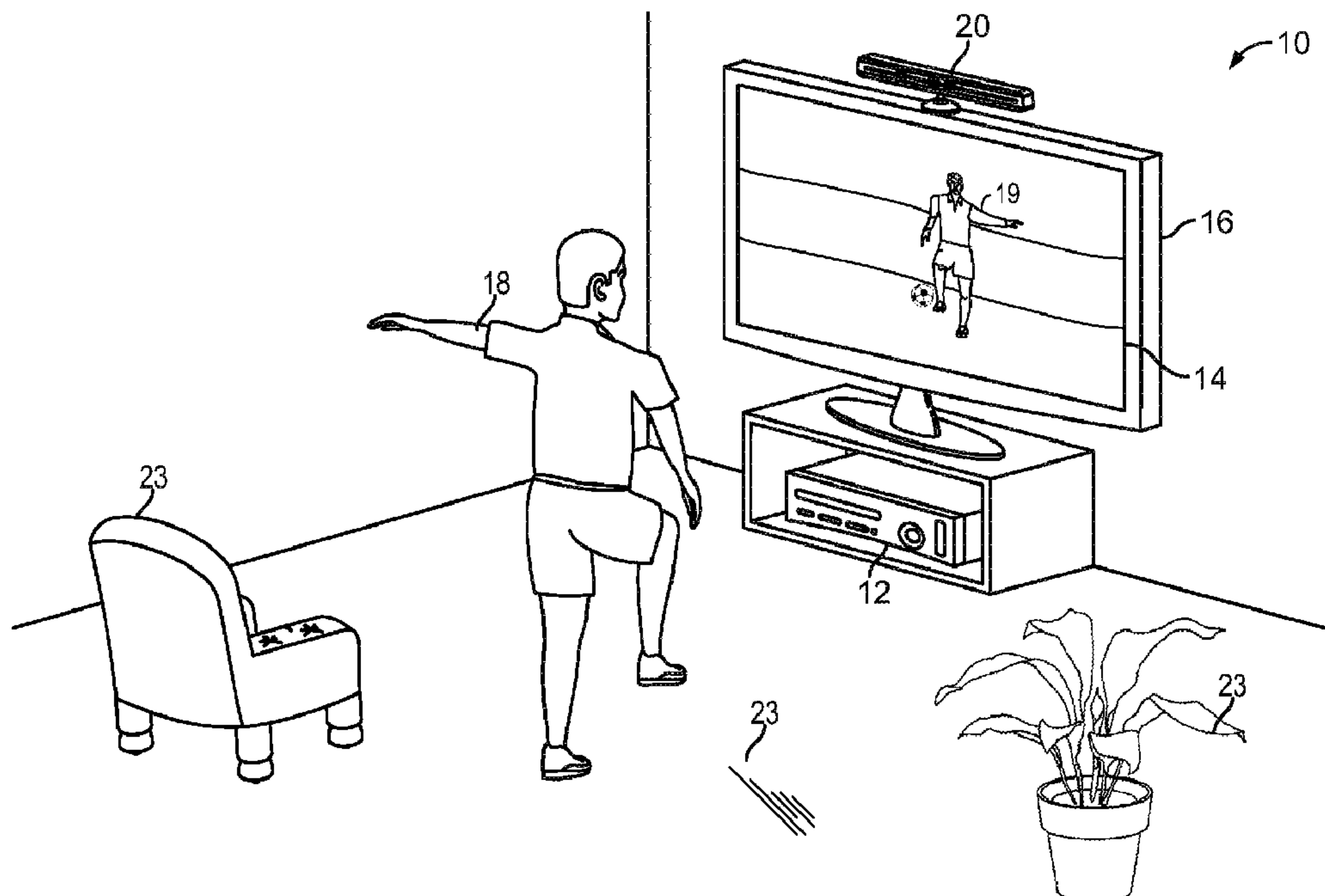
(86) Date de dépôt PCT/PCT Filing Date: 2012/06/04
(87) Date publication PCT/PCT Publication Date: 2012/12/13
(45) Date de délivrance/Issue Date: 2019/09/03
(85) Entrée phase nationale/National Entry: 2013/11/26
(86) N° demande PCT/PCT Application No.: US 2012/040741
(87) N° publication PCT/PCT Publication No.: 2012/170349
(30) Priorités/Priorities: 2011/06/06 (US61/493,850);
2011/10/19 (US13/277,011)

(51) Cl.Int./Int.Cl. *G06F 3/01* (2006.01),
A63F 13/213 (2014.01), *A63F 13/428* (2014.01),
G06F 3/042 (2006.01), *G06F 3/0487* (2013.01)

(72) Inventeurs/Inventors:
AMBRUS, ANTHONY, US;
LEE, KYUNGSUK DAVID, US;
CAMPBELL, ANDREW, US;
HALEY, DAVID, US;
MOUNT, BRIAN, US;
ROBLES, ALBERT, US;
...

(73) Propriétaire/Owner:

(54) Titre : SYSTEME DE RECONNAISSANCE ET DE SUIVI DE DOIGT
(54) Title: SYSTEM FOR FINGER RECOGNITION AND TRACKING



(57) Abrégé/Abstract:

A system and method are disclosed relating to a pipeline for generating a computer model of a target user, including a hand model of the user's hands and fingers, captured by an image sensor in a NUI system. The computer model represents a best estimate of

(72) **Inventeurs(suite)/Inventors(continued):** OSBORN, DANIEL, US; WRIGHT, SHAWN, US; SHARKASI, NAHIL, US; HILL, DAVE, US; MCCULLOCH, DANIEL, US

(73) **Propriétaires(suite)/Owners(continued):** MICROSOFT TECHNOLOGY LICENSING, LLC, US

(74) **Agent:** SMART & BIGGAR

(57) **Abrégé(suite)/Abstract(continued):**

the position and orientation of a user's hand or hands. The generated hand model may be used by a gaming or other application to determine such things as user gestures and control actions.

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

(19) World Intellectual Property
Organization
International Bureau



WIPO | PCT



(10) International Publication Number

WO 2012/170349 A3

(43) International Publication Date
13 December 2012 (13.12.2012)

(51) International Patent Classification:
G06T 7/20 (2006.01)

(21) International Application Number:

PCT/US2012/040741

(22) International Filing Date:

4 June 2012 (04.06.2012)

(25) Filing Language:

English

(26) Publication Language:

English

(30) Priority Data:

61/493,850	6 June 2011 (06.06.2011)	US
13/277,011	19 October 2011 (19.10.2011)	US

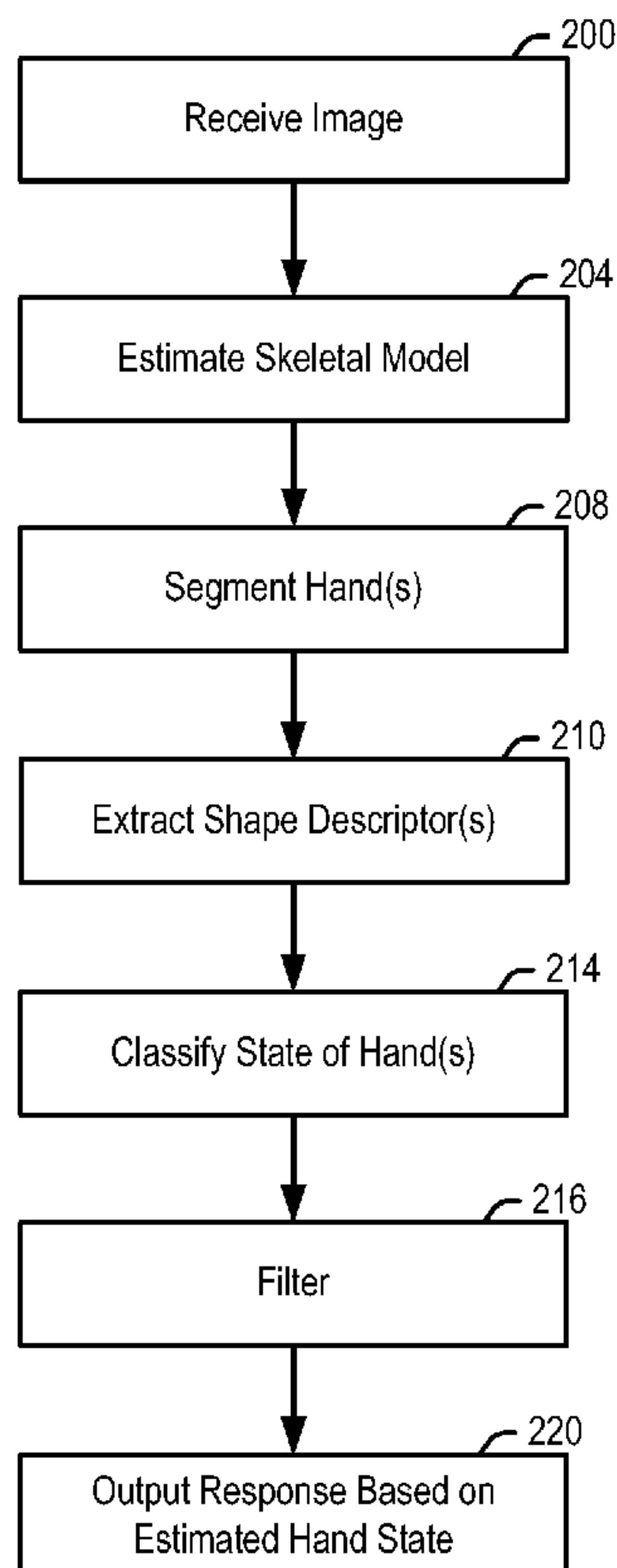
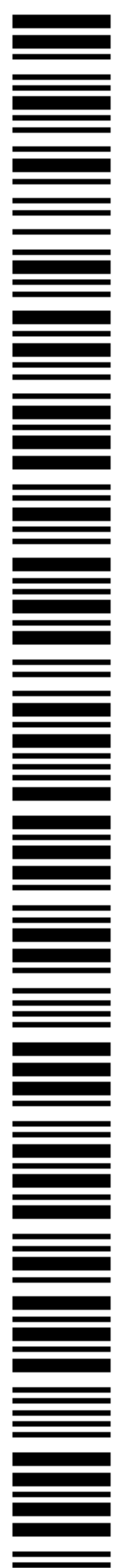
(71) Applicant (for all designated States except US): **MICROSOFT CORPORATION** [US/US]; One Microsoft Way, Redmond, Washington 98052-6399 (US).

(72) Inventors: **AMBRUS, Anthony**; c/o Microsoft Corporation, LCA - International Patents, One Microsoft Way, Redmond, Washington 98052-6399 (US). **LEE, Kyung-suk David**; c/o Microsoft Corporation, LCA - International Patents, One Microsoft Way, Redmond, Washington 98052-6399 (US). **CAMPBELL, Andrew**; c/o Microsoft Corporation, LCA - International Patents, One Microsoft Way, Redmond, Washington 98052-6399 (US). **HALEY, David**; c/o Microsoft Corporation, LCA - International Patents, One Microsoft Way, Redmond, Washington 98052-6399 (US). **MOUNT, Brian**; c/o Microsoft Corporation, LCA - International Patents, One Microsoft Way,

[Continued on next page]

(54) Title: SYSTEM FOR FINGER RECOGNITION AND TRACKING

(57) Abstract: A system and method are disclosed relating to a pipeline for generating a computer model of a target user, including a hand model of the user's hands and fingers, captured by an image sensor in a NUI system. The computer model represents a best estimate of the position and orientation of a user's hand or hands. The generated hand model may be used by a gaming or other application to determine such things as user gestures and control actions.

**Fig. 6****WO 2012/170349 A3**

WO 2012/170349 A3

Redmond, Washington 98052-6399 (US). **ROBLES, Albert**; c/o Microsoft Corporation, LCA - International Patents, One Microsoft Way, Redmond, Washington 98052-6399 (US). **OSBORN, Daniel**; c/o Microsoft Corporation, LCA - International Patents, One Microsoft Way, Redmond, Washington 98052-6399 (US). **WRIGHT, Shawn**; c/o Microsoft Corporation, LCA - International Patents, One Microsoft Way, Redmond, Washington 98052-6399 (US). **SHARKASI, Nahil**; c/o Microsoft Corporation, LCA - International Patents, One Microsoft Way, Redmond, Washington 98052-6399 (US). **HILL, Dave**; c/o Microsoft Corporation, LCA - International Patents, One Microsoft Way, Redmond, Washington 98052-6399 (US). **MCCULLOCH, Daniel**; c/o Microsoft Corporation, LCA - International Patents, One Microsoft Way, Redmond, Washington 98052-6399 (US).

- (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, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, 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, 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, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
- *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

- *with international search report (Art. 21(3))*
- *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

(88) Date of publication of the international search report:

31 January 2013

SYSTEM FOR FINGER RECOGNITION AND TRACKING

BACKGROUND

[0001] In the past, computing applications such as computer games and multimedia applications used controllers, remotes, keyboards, mice, or the like to allow users to
5 manipulate game characters or other aspects of an application. More recently, computer games and multimedia applications have begun employing cameras and software gesture recognition engines to provide a natural user interface (“NUI”). With NUI, raw joint data and user gestures are detected, interpreted and used to control game characters or other aspects of an application.

10 [0002] One of the challenges of a NUI system is distinguishing a person in the field of view of an image sensor, and correctly identifying the positions of his or her body parts including hands and fingers within the field of view. Routines are known for tracking arms, legs, heads and torso. However, given the subtle detail and wide variety of positions of a user’s hands, conventional systems are not able to satisfactorily recognize and track a
15 user’s body including finger and hand positions.

SUMMARY

[0003] Disclosed herein are systems and methods for recognizing and tracking a user’s skeletal joints, including hand and finger positions with a NUI system. In examples, the tracking of hand and finger positions may be used by NUI systems for
20 triggering events such as selecting, engaging, or grabbing and dragging objects on a screen. A variety of other gestures, control actions and applications may be enabled by the present technology for recognizing and tracking hand and finger positions and motions. By determining states of a user’s hand and fingers, interactivity of a user with a NUI system may be increased, and simpler and more intuitive interfaces may be presented to a
25 user.

[0004] In one example, the present disclosure relates to a method for generating a model of a user’s hand including one or more fingers for a natural user interface, comprising: (a) receiving image data of a user interacting with the natural user interface; and (b) analyzing the image data to identify the hand in the image data, said step (b)
30 including the steps of: (b)(1) analyzing depth data from the image data captured in said step (a) to segment the image data into data of the hand, and (b)(2) extracting a shape descriptor by applying one or more filters to the image data of the hand identified in said step (b)(1), the one or more filters analyzing image data of the hand as compared to image

data outside of a boundary of the hand to discern a shape and orientation of the hand.

[0005] In a further example, the present disclosure relates to a system for generating a model of a user's hand including one or more fingers for a natural user interface, the system comprising: a skeletal recognition engine for recognizing a skeleton of a user from received image data; an image segmentation engine for segmenting one or more regions of the body into a region representing a hand of the user; and a descriptor extraction engine for extracting data representative of a hand including one or more fingers and an orientation of the hand, the descriptor extraction engine applying a plurality of filters for analyzing pixels in the region representing the hand, each filter in the plurality of filters determining a position and orientation of the hand, the descriptor extraction engine combining the results of each filter to arrive at a best estimate of the position and orientation of the hand.

[0006] In another example, the present disclosure relates to a computer-readable storage medium not consisting of a modulated data signal, the computer-readable storage medium having computer-executable instructions for programming a processor to perform a method for generating a model of a user's hand including one or more fingers for a natural user interface, the method comprising: (a) receiving image data of a user interacting with the natural user interface; (b) analyzing the image data to identify the hand in the image data; and (c) comparing the image data of the identified hand against predefined hand positions to determine if the user has performed one of the following predefined hand gestures or control actions: (c)(1) counting on the user's fingers, (c)(2) performing an "a-okay" gesture, (c)(3) actuation of a virtual button, (c)(4) pinching together of the thumb and a finger of the hand, (c)(5) writing or drawing, (c)(6) sculpting, (c)(7) puppeteering, (c)(8) turning a knob or combination lock, (c)(9) shooting a gun, (c)(10) performing a flicking gesture, (c)(11) performing a gesture where a finger can be used on an open palm to scroll across and navigate through the virtual space, and (c)(12) moving fingers in a scissor motion to control the legs of a virtual character.

[0006a] According to another aspect of the present invention, there is provided a method for generating a model of a user's hand including one or more fingers, comprising: (a) receiving position data representing a position of a user interacting with a sensor, the position data including at least one of depth and image data representing the user's hand; and (b) analyzing the position data to identify the user's hand in the position data, said step (b) including the steps of: (b)(1) analyzing depth data from the position data captured in said step (a) to segment the position data into data of the hand, and (b)(2) extracting a set of feature descriptors by applying one or more filters to the image data of the hand identified in said step (b)(1), the one or more filters analyzing image data of the hand as

compared to image data outside of a boundary of the hand to discern features of the hand including a shape and orientation of the hand.

5 [0006b] According to still another aspect of the present invention, there is provided a system for generating a model of a user's hand including one or more fingers, the system including a sensing mechanism operatively coupled to a computing device, the system comprising: a skeletal recognition engine for recognizing at least a portion of a skeleton of a user from received data including at least one of image and depth data; an image segmentation engine for segmenting one or more regions of the body into a region representing a hand of the user; and a descriptor extraction engine for extracting data representative of a hand including one or more fingers and an orientation of the hand, the
10 descriptor extraction engine applying a plurality of filters for analyzing pixels in the region representing the hand, each filter in the plurality of filters determining a position and orientation of the hand, the descriptor extraction engine combining the results of each filter to arrive at a best estimate of the position and orientation of the hand.

15 [0006c] According to yet another aspect of the present invention, there is provided a computer-readable medium having stored thereon computer-executable instructions, that when executed, perform a method as described above or detailed below.

[0007] This Summary is provided to introduce a selection of concepts in a simplified form that are further described below in the Detailed Description. This Summary is not intended to identify key features or essential features of the claimed subject matter, nor is it intended to be used as an aid in
20 determining the scope of the claimed subject matter. Furthermore, the claimed subject matter is not limited to implementations that solve any or all disadvantages noted in any part of this disclosure.

BRIEF DESCRIPTION OF THE DRAWINGS

- [0008] FIGURE 1A illustrates an example embodiment of a target recognition, analysis, and tracking system.
- [0009] FIGURE 1B illustrates a further example embodiment of a target recognition, analysis, and tracking system.
- [0010] FIGURE 1C illustrates a still further example embodiment of a target recognition, analysis, and tracking system.
- [0011] FIGURE 2 illustrates an example embodiment of a capture device that may be used in a target recognition, analysis, and tracking system.
- [0012] FIGURE 3 shows an exemplary body model used to represent a human target.
- [0013] FIGURE 4 shows a substantially frontal view of an exemplary skeletal model used to represent a human target.
- [0014] FIGURE 5 shows a skewed view of an exemplary skeletal model used to represent a human target.
- [0015] FIGURE 6 shows a flowchart of a pipeline for tracking a target according to an embodiment of the present technology.
- [0016] FIGURE 7 illustrates an example method of determining a state of a hand of a user in accordance with an embodiment of the present disclosure.
- [0017] FIGURE 8 is a flowchart of the operation of an image segmentation engine according to an embodiment of the present disclosure.
- [0018] FIGURE 9 is a flowchart of the operation of a pixel classification filter according to an embodiment of the present disclosure.
- [0019] FIGURE 10 is a decision tree of the operation of a pixel classification filter according to an embodiment of the present disclosure.
- [0020] FIGURES 11A and 11B illustrate fingertip identification using a pixel classification filter according to an embodiment of the present disclosure.
- [0021] FIGURE 12 illustrates finger identification using a pixel classification filter according to an embodiment of the present disclosure.
- [0022] FIGURE 13 illustrates a portion of a hand identified using a pixel classification filter according to an embodiment of the present disclosure.
- [0023] FIGURE 14 illustrates hand and finger identification using a pixel classification filter according to an embodiment of the present disclosure.

[0024] FIGURE 15 is a flowchart of the operation of a curvature analysis filter according to an embodiment of the present disclosure.

[0025] FIGURE 16 illustrates hand and finger identification using a curvature analysis filter according to an embodiment of the present disclosure.

5 [0026] FIGURE 17 illustrates open and closed hand analysis using a depth histogram filter according to an embodiment of the present disclosure.

[0027] FIGURE 18 is a flow diagram of a supervisor filter for classifying a hand position based on hand filters.

[0028] FIGURE 19A illustrates an example embodiment of a computing environment
10 that may be used to interpret one or more gestures in a target recognition, analysis, and tracking system.

[0029] FIGURE 19B illustrates another example embodiment of a computing environment that may be used to interpret one or more gestures in a target recognition, analysis, and tracking system.

15 **DETAILED DESCRIPTION**

[0030] Embodiments of the present technology will now be described with reference to Figs. 1A - 19B which in general relate to a pipeline for generating a computer model of a target user, including a hand model of the user's hands and fingers, captured by an image sensor in a NUI system. The computer model may be generated once per frame of
20 captured image data, and represents a best estimate of the position, including pose, of a user during the captured frame. The generated hand model for each frame may be used by a gaming or other application to determine such things as user gestures and control actions. The hand model may also be fed back into the pipeline to aid in future model determinations.

25 [0031] Referring initially to Figs. 1A-2, the hardware for implementing the present technology includes a target recognition, analysis, and tracking system 10 which may be used to recognize, analyze, and/or track a human target such as the user 18. Embodiments of the target recognition, analysis, and tracking system 10 include a computing environment 12 for executing a gaming or other application. The computing environment
30 12 may include hardware components and/or software components such that computing environment 12 may be used to execute applications such as gaming and non-gaming applications. In one embodiment, computing environment 12 may include a processor such as a standardized processor, a specialized processor, a microprocessor, or the like that

may execute instructions stored on a processor readable storage device for performing processes described herein.

[0032] The system 10 further includes a capture device 20 for capturing image and audio data relating to one or more users and/or objects sensed by the capture device. In
5 embodiments, the capture device 20 may be used to capture information relating to body and hand movements and/or gestures and speech of one or more users, which information is received by the computing environment and used to render, interact with and/or control aspects of a gaming or other application. Examples of the computing environment 12 and capture device 20 are explained in greater detail below.

10 [0033] Embodiments of the target recognition, analysis and tracking system 10 may be connected to an audio/visual (A/V) device 16 having a display 14. The device 16 may for example be a television, a phone, a monitor for a computer, a high-definition television (HDTV), or the like that may provide game or application visuals and/or audio to a user. For example, the computing environment 12 may include a video adapter such as a
15 graphics card and/or an audio adapter such as a sound card that may provide audio/visual signals associated with the game or other application. The A/V device 16 may receive the audio/visual signals from the computing environment 12 and may then output the game or application visuals and/or audio associated with the audio/visual signals to the user 18. According to one embodiment, the audio/visual device 16 may be connected to the
20 computing environment 12 via, for example, an S-Video cable, a coaxial cable, an HDMI cable, a DVI cable, a VGA cable, a component video cable, or the like.

[0034] In embodiments, the computing environment 12, the A/V device 16 and the capture device 20 may cooperate to render an avatar or on-screen character 19 on display 14. For example, Fig. 1A shows a user 18 playing a soccer gaming application. The user's
25 movements are tracked and used to animate the movements of the avatar 19. In embodiments, the avatar 19 mimics the movements of the user 18 in real world space so that the user 18 may perform movements and gestures which control the movements and actions of the avatar 19 on the display 14.

[0035] As explained above, motion estimation routines such as skeleton mapping
30 systems may lack the ability to detect subtle gestures of a user, such as for example the movement of a user's hand. For example, a user may wish to interact with NUI system 10 by scrolling through and controlling a user interface 21 with his hand as shown in Fig. 1B. A user may alternatively attempt to perform various gestures, such as for example by opening and/or closing her hand as shown as 23 and 25 in Fig. 1C.

[0036] Accordingly, systems and methods, described below herein, are directed to determining a state of a hand of a user. For example, the action of closing and opening the hand may be used by such systems for triggering events such as selecting, engaging, or grabbing and dragging objects, e.g., object 27 (Fig. 1C), on the screen. These actions
 5 otherwise would correspond to pressing a button when using a controller. Such refined controller-free interaction can be used as an alternative to approaches based on hand waving or hovering, which may be unintuitive or cumbersome. A variety of other gestures, control actions and applications may be enabled by the present technology for recognizing and tracking hand motions, some of which are described in further detail
 10 below. By determining states of a user's hand as described below, interactivity of a user with the system may be increased and simpler and more intuitive interfaces may be presented to a user.

[0037] Figs. 1A-1B include static, background objects 23, such as a floor, chair and plant. These are objects within the field of view (FOV) captured by capture device 20, but
 15 do not change from frame to frame. In addition to the floor, chair and plant shown, static objects may be any objects picked up by the image cameras in capture device 20. The additional static objects within the scene may include any walls, ceiling, windows, doors, wall decorations, etc.

[0038] Suitable examples of a system 10 and components thereof are found in the
 20 following co-pending patent applications:
 United States Patent Application Serial No. 12/475,094, entitled
 "Environment and/or Target Segmentation," filed May 29, 2009; United States Patent
 Application Serial No. 12/511,850, entitled "Auto Generating a Visual Representation,"
 filed July 29, 2009; United States Patent Application Serial No. 12/474,655, entitled
 25 "Gesture Tool," filed May 29, 2009; United States Patent Application Serial No.
 12/603,437, entitled "Pose Tracking Pipeline," filed October 21, 2009; United States
 Patent Application Serial No. 12/475,308, entitled "Device for Identifying and Tracking
 Multiple Humans Over Time," filed May 29, 2009, United States Patent Application Serial
 No. 12/575,388, entitled "Human Tracking System," filed October 7, 2009; United States
 30 Patent Application Serial No. 12/422,661, entitled "Gesture Recognizer System
 Architecture," filed April 13, 2009; and United States Patent Application Serial No.
 12/391,150, entitled "Standard Gestures," filed February 23, 2009.

[0039] Fig. 2 illustrates an example embodiment of the capture device 20 that may be used in the target recognition, analysis, and tracking system 10. In an example embodiment, the capture device 20 may be configured to capture video having a depth image that may include depth values via any suitable technique including, for example, time-of-flight, structured light, stereo image, or the like. According to one embodiment, the capture device 20 may organize the calculated depth information into “Z layers,” or layers that may be perpendicular to a Z axis extending from the depth camera along its line of sight. X and Y axes may be defined as being perpendicular to the Z axis. The Y axis may be vertical and the X axis may be horizontal. Together, the X, Y and Z axes define the 3-D real world space captured by capture device 20.

[0040] As shown in Fig. 2, the capture device 20 may include an image camera component 22. According to an example embodiment, the image camera component 22 may be a depth camera that may capture the depth image of a scene. The depth image may include a two-dimensional (2-D) pixel area of the captured scene where each pixel in the 2-D pixel area may represent a depth value such as a length or distance in, for example, centimeters, millimeters, or the like of an object in the captured scene from the camera.

[0041] As shown in Fig. 2, according to an example embodiment, the image camera component 22 may include an IR light component 24, a three-dimensional (3-D) camera 26, and an RGB camera 28 that may be used to capture the depth image of a scene. For example, in time-of-flight analysis, the IR light component 24 of the capture device 20 may emit an infrared light onto the scene and may then use sensors (not shown) to detect the backscattered light from the surface of one or more targets and objects in the scene using, for example, the 3-D camera 26 and/or the RGB camera 28.

[0042] In some embodiments, pulsed infrared light may be used such that the time between an outgoing light pulse and a corresponding incoming light pulse may be measured and used to determine a physical distance from the capture device 20 to a particular location on the targets or objects in the scene. Additionally, in other example embodiments, the phase of the outgoing light wave may be compared to the phase of the incoming light wave to determine a phase shift. The phase shift may then be used to determine a physical distance from the capture device 20 to a particular location on the targets or objects.

[0043] According to another example embodiment, time-of-flight analysis may be used to indirectly determine a physical distance from the capture device 20 to a particular location on the targets or objects by analyzing the intensity of the reflected beam of light over time via various techniques including, for example, shuttered light pulse imaging.

5 [0044] In another example embodiment, the capture device 20 may use a structured light to capture depth information. In such an analysis, patterned light (*i.e.*, light displayed as a known pattern such as a grid pattern or a stripe pattern) may be projected onto the scene via, for example, the IR light component 24. Upon striking the surface of one or more targets or objects in the scene, the pattern may become deformed in response. Such
10 a deformation of the pattern may be captured by, for example, the 3-D camera 26 and/or the RGB camera 28 and may then be analyzed to determine a physical distance from the capture device 20 to a particular location on the targets or objects.

[0045] According to another embodiment, the capture device 20 may include two or more physically separated cameras that may view a scene from different angles, to obtain
15 visual stereo data that may be resolved to generate depth information. In another example embodiment, the capture device 20 may use point cloud data and target digitization techniques to detect features of the user. Other sensor systems may be used in further embodiments, such as for example an ultrasonic system capable of detecting x, y and z axes.

20 [0046] The capture device 20 may further include a microphone 30. The microphone 30 may include a transducer or sensor that may receive and convert sound into an electrical signal. According to one embodiment, the microphone 30 may be used to reduce feedback between the capture device 20 and the computing environment 12 in the target recognition, analysis, and tracking system 10. Additionally, the microphone 30 may
25 be used to receive audio signals that may also be provided by the user to control applications such as game applications, non-game applications, or the like that may be executed by the computing environment 12.

[0047] In an example embodiment, the capture device 20 may further include a processor 32 that may be in operative communication with the image camera component
30 22. The processor 32 may include a standardized processor, a specialized processor, a microprocessor, or the like that may execute instructions that may include instructions for receiving the depth image, determining whether a suitable target may be included in the depth image, converting the suitable target into a skeletal representation or model of the target, or any other suitable instruction.

[0048] The capture device 20 may further include a memory component 34 that may store the instructions that may be executed by the processor 32, images or frames of images captured by the 3-D camera or RGB camera, or any other suitable information, images, or the like. According to an example embodiment, the memory component 34
5 may include random access memory (RAM), read only memory (ROM), cache, Flash memory, a hard disk, or any other suitable storage component. As shown in Fig. 2, in one embodiment, the memory component 34 may be a separate component in communication with the image camera component 22 and the processor 32. According to another embodiment, the memory component 34 may be integrated into the processor 32 and/or
10 the image camera component 22.

[0049] As shown in Fig. 2, the capture device 20 may be in communication with the computing environment 12 via a communication link 36. The communication link 36 may be a wired connection including, for example, a USB connection, a Firewire connection, an Ethernet cable connection, or the like and/or a wireless connection such as a wireless
15 802.11b, g, a, or n connection. According to one embodiment, the computing environment 12 may provide a clock to the capture device 20 that may be used to determine when to capture, for example, a scene via the communication link 36.

[0050] Additionally, the capture device 20 may provide the depth information and images captured by, for example, the 3-D camera 26 and/or the RGB camera 28. With the
20 aid of these devices, a partial skeletal model may be developed in accordance with the present technology, with the resulting data provided to the computing environment 12 via the communication link 36.

[0051] The computing environment 12 may further include a gesture recognition engine 190 for recognizing gestures as explained below. In accordance with the present
25 system, the computing environment 12 may further include a skeletal recognition engine 192, an image segmentation engine 194, a descriptor extraction engine 196 and a classifier engine 198. Each of these software engines is described in greater detail below.

[0052] Fig. 3 shows a non-limiting visual representation of an example body model
70 generated by skeletal recognition engine 192. Body model 70 is a machine
30 representation of a modeled target (*e.g.*, user 18 from Figs. 1A and 1B). The body model may include one or more data structures that include a set of variables that collectively define the modeled target in the language of a game or other application/operating system.

[0053] A model of a target can be variously configured without departing from the scope of this disclosure. In some examples, a model may include one or more data structures that represent a target as a three-dimensional model including rigid and/or deformable shapes, or body parts. Each body part may be characterized as a mathematical primitive, examples of which include, but are not limited to, spheres, anisotropically-scaled spheres, cylinders, anisotropic cylinders, smooth cylinders, boxes, beveled boxes, prisms, and the like.

[0054] For example, body model 70 of Fig. 3 includes body parts bp1 through bp14, each of which represents a different portion of the modeled target. Each body part is a three-dimensional shape. For example, bp3 is a rectangular prism that represents the left hand of a modeled target, and bp5 is an octagonal prism that represents the left upper-arm of the modeled target. Body model 70 is exemplary in that a body model may contain any number of body parts, each of which may be any machine-understandable representation of the corresponding part of the modeled target.

[0055] A model including two or more body parts may also include one or more joints. Each joint may allow one or more body parts to move relative to one or more other body parts. For example, a model representing a human target may include a plurality of rigid and/or deformable body parts, wherein some body parts may represent a corresponding anatomical body part of the human target. Further, each body part of the model may include one or more structural members (*i.e.*, “bones” or skeletal parts), with joints located at the intersection of adjacent bones. It is to be understood that some bones may correspond to anatomical bones in a human target and/or some bones may not have corresponding anatomical bones in the human target.

[0056] The bones and joints may collectively make up a skeletal model, which may be a constituent element of the body model. In some embodiments, a skeletal model may be used instead of another type of model, such as model 70 of Fig. 3. The skeletal model may include one or more skeletal members for each body part and a joint between adjacent skeletal members. Exemplary skeletal model 80 and exemplary skeletal model 82 are shown in Figs. 4 and 5, respectively. Fig. 4 shows a skeletal model 80 as viewed from the front, with joints j1 through j33. Fig. 5 shows a skeletal model 82 as viewed from a skewed view, also with joints j1 through j33. A skeletal model may include more or fewer joints without departing from the spirit of this disclosure. Further embodiments of the present system explained hereinafter operate using a skeletal model having 31 joints.

[0057] The above described body part models and skeletal models are non-limiting examples of types of models that may be used as machine representations of a modeled target. Other models are also within the scope of this disclosure. For example, some models may include polygonal meshes, patches, non-uniform rational B-splines, subdivision surfaces, or other high-order surfaces. A model may also include surface textures and/or other information to more accurately represent clothing, hair, and/or other aspects of a modeled target. A model may optionally include information pertaining to a current pose, one or more past poses, and/or model physics. It is to be understood that a variety of different models that can be posed are compatible with the herein described target recognition, analysis, and tracking system.

[0058] Software pipelines for generating skeletal models of one or more users within a FOV of capture device 20 are known. One such system is disclosed for example in United States Patent Application Serial No. 12/876,418, entitled "System For Fast, Probabilistic Skeletal Tracking," filed September 7, 2010.

Under certain conditions, for example where a user is sufficiently close to capture device 20 and at least one of the user's hands is distinguishable from other background noise, a software pipeline may further be able to generate hand models for the hand and/or fingers of one or more users within the FOV.

[0059] Fig. 6 is a flowchart of a software pipeline for recognizing and tracking a user's hand and/or fingers. In step 200, the pipeline receives a depth image from capture device 20. A depth image of a portion of a user is illustrated in Fig. 7 at 302. Each pixel in the depth image includes depth information, e.g., as illustrated in Fig. 7 by a gray-scale gradient. For example, at 302, the left hand of the user is closer to the capture device 20, as indicated by the darker region of the left hand. The capture device or depth camera captures images of a user within an observed scene. As described below, a depth image of a user may be used to determine distance information of the regions of the user, scale information of the user, curvature, and skeletal information of the user.

[0060] In step 204, the skeletal recognition engine 192 of the pipeline estimates a skeleton model of the user as described above to obtain a virtual skeleton from a depth image obtained in step 200. For example, in Fig. 7, a virtual skeleton 304 is shown as estimated from the depth image shown at 302 of the user.

[0061] In step 208, the pipeline segments a hand or hands of the user via the image segmentation engine 194 of the pipeline. In some examples, image segmentation engine 194 may additionally segment one or more regions of the body in addition to the hands.

Segmenting a hand of a user includes identifying a region of the depth image corresponding to the hand, where the identifying is at least partially based on the skeleton information obtained in step 204. Fig. 7 illustrates an example of segmentation of the depth image of a user into different regions 306 based on estimated skeleton 304, as indicated by the differently shaded regions. Fig. 7 shows localized hand region 308 corresponding to the user's raised right hand.

[0062] Hands or body regions may be segmented or localized in a variety of ways and may be based on selected joints identified in the skeleton estimation described above. As one example, hand detection and localization in the depth image may be based on the estimated wrist and/or hand tip joints from the estimated skeleton. For example, in some embodiments, hand segmentation in the depth image may be performed using a topographical search of the depth image around the hand joints, locating nearby local extrema in the depth image as candidates for finger tips. The image segmentation engine 194 then segments the rest of the hand by taking into account a body size scaling factor as determined from the estimated skeleton, as well as depth discontinuities for boundary identification.

[0063] As another example, a flood-fill approach may be employed to identify regions of the depth image corresponding to a user's hands. In a flood-fill approach, the depth image may be searched from a starting point and a starting direction, e.g., the starting point may be the wrist joint and the starting direction may be a direction from the elbow to the wrist joint. Nearby pixels in the depth image may be iteratively scored based on the projection on the starting direction as a way for giving preference to points moving away from the elbow and toward the hand tip, while depth consistency constraints such as depth discontinuities may be used to identify boundaries or extreme values of a user's hands in the depth image. In some examples, threshold distance values may be used to limit the depth map search in both the positive and negative directions of the starting direction based on fixed values or scaled based on an estimated size of the user, for example.

[0064] As still another example, a bounding sphere or other suitable bounding shape, positioned based on skeleton joints (e.g. wrist or hand tip joints), may be used to include all pixels in the depth image up to a depth discontinuity. For example, a window may be slid over the bounding sphere to identify depth discontinuities which may be used to establish a boundary in the hand region of the depth image.

[0065] The bounding shape method may also be used to place a bounding shape around a center of the palm of the hand which may be iteratively identified. One example of such a iterative bounding method is disclosed in a presentation by David Tuft, titled "Kinect Developer Summit at GDC 2011: Kinect for XBOX 360," attached hereto as Attachment 1, and in a publication by K. Abe, H. Saito, S. Ozawa, titled "*3D drawing system via hand motion recognition from cameras*", IEEE International Conference on Systems, Man, and Cybernetics, vol. 2, 2000.

[0066] In general, such a method involves several iterative passes to cull pixels from the model. In each pass, the method culls pixels outside the sphere or other shape centered at the hand. Next, the method culls pixels too far from the tip of the hand (along the arm vector). Then the method performs an edge detection step to edge detect hand boundary and remove unconnected islands. Example steps from such a method are shown in the flowchart of Fig. 8. In step 224, a bounding shape is generated around a center of the hand given by the hand joint data from the skeletal recognition engine 192. The bounding shape is large enough to encompass the entire hand and is three-dimensional. In step 226, the pixels outside the bounding shape are culled.

[0067] It may happen that a user's hand is close to his or her body, or to the user's second hand, in the depth image, and the data from those other body portions will initially be included in the segmented image. Connected component labeling may be performed to label different centroids in the segmented image. The centroid, which is most likely the hand, is selected, based on its size and the location of the hand joint. The centroids not selected may be culled. In step 230, pixels which are too far from the tip of the hand along a vector from the attached arm may also be culled.

[0068] The skeletal data from the skeletal recognition engine 192 may be noisy, so the data for the hand is further refined to identify the center of the hand. This may be done by iterating over the image and measuring the distance of each pixel to the edge of the silhouette of the hand. The image segmentation engine 194 may then perform a weighted average to figure out the maximum/minimum distance. That is, in step 232, for each pixel in the segmented hand image, a maximum distance along the x and y axes to an edge of the hand silhouette is identified, and a minimum distance along the x and y axes to an edge of the hand silhouette is identified. The distance to the edge is taken as a weight, and a weighted average of the minimum determined distances is then taken across all measured pixels to figure out the likely center of the hand position within the image (step 234).

Using the new center, the process may be iteratively repeated until the change in the palm center from the previous iteration is within some tolerance.

[0069] In some approaches, segmenting of hand regions may be performed when a user raises the hand outward or above or in front of the torso. In this way, identification of hand regions in the depth image may be less ambiguous since the hand regions may be distinguished from the body more easily. Hand images are particularly clear when a user's hand is oriented palm toward the capture device 20, at which point, the features of that hand can be detected as a silhouette. Features may be noisy, but a silhouetted hand allows some informed decisions about what a hand is doing, based on for example detecting gaps between fingers and seeing the overall shape of the hand and mapping that using a variety of different approaches. Detecting those gaps and other features allows recognition of particular fingers, and a general direction of where that finger is pointing.

[0070] It should be understood that the example hand segmentation examples described above are presented for the purpose of example and are not intended to limit the scope of this disclosure. In general, any hand or body part segmentation method may be used alone or in combination with each other and/or one of the example methods described above.

[0071] Continuing with the pipeline of Fig. 7, step 210 involves extracting a shape descriptor for the region, e.g., the region of the depth image corresponding to a hand as identified in step 208. The shape descriptor in step 210 is extracted by the descriptor extraction engine 196, and may be any suitable representation of the hand region which is used to classify the hand region. In some embodiments, the shape descriptor may be a vector or a set of numbers used to codify or describe the shape of the hand region.

[0072] The descriptor extraction engine 196 may use any of a variety of filters in step 210 to extract a shape descriptor. One filter may be referred to as a pixel classifier, which will now be described with reference to the flowchart of Fig. 9, the decision tree of Fig. 10 and the illustrations of Figs. 11-14. In step 240, a pixel in the foreground of the segmented image is selected. These are the pixels at least nominally believed to be part of the user's hand. A box of predefined size is taken around the selected pixel, with the selected pixel at the center. In embodiments, the size of the box may be selected to be 1.5 times the width of a normalized finger. A "normalized finger" is the finger of the user which has been adjusted to a normalized size based on the size of the skeletal model and a detected distance of the user from the capture device 20. The following steps are run successively for each pixel nominally believed to be part of the hand.

[0073] In step 242, the pixel classifier filter determines how many edges of the box are intersected. An intersection is where the image transitions from a foreground (on the hand) to a background (not on the hand). For example, Fig. 11A shows a finger 276, a selected pixel 278 on the finger, and the above-described box 280 around the pixel at a radius r . The box is intersected at two points along a single edge; at points 281a and 281b. The points 281a, 281b are where the image transitions from the foreground (finger) to the background. All pixels 278 having two intersection points with edges of their respective boxes 280 are considered finger tips (or part of the knuckle or arm as explained below) for the purposes of defining hand centroids as explained below.

10 [0074] In step 246, the pixel classifier filter determines whether the intersections are in the same or different edges. As seen in Fig. 11B, a finger may intersect the box 280 along two adjacent edges instead of along the same edge. This information will be used to determine the direction at which the finger is pointed as explained below.

[0075] As opposed to a fingertip, a pixel which intersects its box 280 at four points will be considered a finger for the purposes of defining hand centroids as explained below. For example, Fig. 12 shows an example where the selected pixel 278 is sufficiently distal from the fingertip, that there are four intersection points 281a, 281b, 281c and 281d with the box 280.

[0076] In step 242 of the flowchart of Fig. 9, and at 264 of the decision tree of Fig. 10, the pixel classifier filter checks how many edges of box 280 are intersected. If no edges are intersected, the selected pixel is considered to be within the user's palm at 265. That is, because the size of the box 280 is selected so that at least two edges will be intersected if the pixel lies in a finger or fingertip, if the pixel lies on the hand, and no edges are intersected, the pixel is considered to lie in the palm. If two edges are intersected, the filter goes to 266 to check whether the corners of non-intersected edges are solid (on the hand) or empty (background) as explained below. If four edges are intersected in 267, it is considered a finger as explained above. If the edges of a box 280 are intersected six times in 268, this is considered an invalid reading and is discarded (step 250).

30 [0077] Referring again to 266, where two edges are intersected, it could be a fingertip, but it could also be a space between two adjacent fingers. The pixel classifier filter therefore checks the corners of the non-intersected edges (step 248). Where the corners of non-intersected edges are solid, this means the box lies on the hand at those corners and the intersection points define a valley between adjacent fingers. Conversely,

where the corners of non-intersected edges are empty (as shown in the illustration associated with 266), this means the box lies on background pixels at those corners and the intersection points define a part of the hand.

[0078] If the corners are empty, the pixel classifier filter checks at 269 whether the distance, referred to as chord length, between the intersection points is less than the maximum width of a finger (step 252). That is, where there are two intersection points, it could be a fingertip as shown in Fig. 11A. However, the pixel could also be a part of the arm or part of the hand, such as the knuckle, as shown in Fig. 13. If so, the length of chord 282 could be greater than the maximum width of a finger. If so, the pixel 278 for which box 280 is being examined is said to lie on the arm or knuckle at 271 (Fig. 10).

[0079] In addition to identifying a fingertip or finger, a two-point or a four-point intersection also can reveal a direction in which the fingertip/finger is pointing. For example, in Fig. 11A, there were two intersections less than the maximum width of a finger, so it was determined that the pixel 278 lies in a fingertip. However, given such an intersection, inferences can be drawn about the direction in which the fingertip is pointing. The same can be said for the finger shown in Fig. 12. Fig. 11A shows the finger 276 pointing straight up. But the fingertip 276 could be pointing in other upward directions as well. Information from other points near the point 278 in fingertip 276 may be used to draw further inferences about the direction.

[0080] Fig. 11B shows a two-point intersection, which provides additional inferences about the direction the finger/fingertip is pointing. That is, the direction may be inferred from the ratio of the distances to the shared corner. Stated another way, the chord length between the points 281a and 281b defines the hypotenuse of a triangle also including the sides between the points 281a, 281b and shared corner. It may be inferred that the finger is pointing in a direction perpendicular to the hypotenuse.

[0081] It may happen that a hand is held with two fingers together, three fingers together, or four fingers together. Thus, after the above steps are run using box 280 for each pixel in the hand, the process may be repeated using a box 280 that is slightly larger than the maximum width of two fingers together, and then repeated again using a box 280 that is slightly larger than the maximum width of three fingers together, etc.

[0082] Once the pixel classifier filter data is gathered, the pixel classifier filter next attempts to construct a hand model from the data in step 258 (Fig. 9). There are small identified regions, or centroids, such as for example regions that are a fingertip, and a region that is a palm, and an idea of the center of the palm from the hand segmentation

step. The classifier engine 198 then examines finger centroids not classified as fingertips but, due to the fact that they intersected at four points, are classified as fingers. The directional orientation has also been identified for finger and fingertip regions. If a finger centroid aligns with a fingertip centroid, and they are in the correct relative location to each other, the algorithm connects those centroids as belonging to the same finger.

5 [0083] Next, the orientation of the finger region is used to project where the knuckle of that finger is believed to be, based on the skeleton size and how big a finger is believed to be. The size, position and orientations of any identified valleys between fingers can also be used to confirm the determined hand model. Next, the projected knuckle position is connected to the palm. Upon completion, the pixel classifier engine determines a skeleton hand model 284, two examples of which are shown in Fig. 14. The model, which may be referred to as a “reduced skeletal model of the hand tracking segments related to finger tips, knuckles connecting the hand and finger, and a center bone to the palm,” includes fingertip centroids connected to finger centroids, connected to knuckle centroids
10 connected to a palm centroid. Data regarding the known geometry and possible positions of a hand from the known arm position can also be used to verify or dispute the determined positions of the fingertip, finger, knuckle and/or palm centroid positions, as well as to discard centroid data that may be determined to not form part of a hand.

[0084] The above will construct a hand model even if one or more portions of the hand are missing from the model. For example, a finger may have been occluded, or too close to the user’s body or other hand to be detected. Or the user may be missing a finger. The pixel classification filter will construct a hand model using the finger and hand positions that it detects.

[0085] Another filter which may be run in addition to or instead of the pixel classification filter may be referred to as a curvature analysis filter. This filter focuses on the curvature along the boundaries of the segmented hand silhouette to determine peaks and valleys in an attempt to differentiate fingers. Referring to the flowchart in Fig. 15, in step 286, starting with a first pixel, the eight surrounding pixels are examined to determine which is the next pixel on the hand. Each pixel is thus assigned a value between 0 to 7 for the connectivity between that pixel and the next. A chain of these numbers is built around the silhouette of the hand which gives you the hand boundaries. These values may be converted into angles and contours around the hand in step 288 to provide a graph of hand contour and peaks, such as shown in Fig. 16. These steps for generating hand contours and peaks are described for example in a paper by F. Leymarie, M.D. Levine, titled
25
30

"Curvature morphology", Computer Vision and Robotics Laboratory, McGill University, Montreal, Quebec, Canada, 1988.

- [0086] Peaks around the hand silhouette are identified in step 289, and each is
 5 analyzed with respect to various features of the peak. A peak may be defined by a start point, a peak and an end point. These three points may form a triangle as explained below. The various features of a peak that may be examined include for example:
- width of a peak;
 - max height of a given peak;
 - 10 • average height of curvature samples within a peak;
 - peak shape ratio (max height/average height);
 - area of a peak;
 - hand to peak distance;
 - elbow to hand direction (x, y and z);
 - 15 • cross product of peak direction and arm direction (how small the angle is between the arm direction and peak direction); and
 - cross product of the vector between the peak start point and max point, and the vector between the max point and the end point.

- [0087] This information may be run through various machine learning techniques in
 20 step 290, such as for example a support vector machine, to differentiate fingers and the hand. Support vector machines are known and described for example in C. Cortes and V. Vapnik, titled *Support-Vector Networks*, Machine Learning, 20(3):273-297, September 1995, and Vladimir N. Vapnik, titled *The Nature of Statistical Learning Theory*. Springer, New York, 1995. In embodiments, noisy data may be smoothed using a Hidden Markov
 25 Model to maintain the state of the hands and filter out noise.

- [0088] The above-described filters may be referred to as silhouette filters in that they
 examine the data relating to the silhouette of a hand. A further filter which may be used is a histogram filter and is referred to as a depth filter in that it uses depth data to construct a
 30 hand model. This filter may be used in addition to or instead of the above-described filters, and may be particularly useful when a user has his or hand pointed toward the image capture device 20.

[0089] In the histogram filter, a histogram of distances in the hand region may be constructed. For example, such a histogram may include fifteen bins, where each bin includes the number of points in the hand region whose distance in the Z-direction (depth) from the closest point to the camera is within a certain distance range associated with that bin. For example, the first bin in such a histogram may include the number of points in the hand region whose distance to the hand centroid is between 0 and 0.40 centimeters, the second bin includes the number of points in the hand region whose distance to the hand centroid is between 0.40 and 0.80 centimeters, and so forth. In this way, a vector may be constructed to codify the shape of the hand. Such vectors may further be normalized based on estimated body size, for example.

[0090] In another example approach, a histogram may be constructed based on distances and/or angles from points in the hand region to a joint, bone segment or palm plane from the user's estimated skeleton, e.g., the elbow joint, wrist joint, etc. Fig. 17 illustrates two graphs indicative of the histograms determined for a closed hand and an open hand.

[0091] It should be understood that the filter examples of shape descriptors are exemplary in nature and are not intended to limit the scope of this disclosure. In general, any suitable shape descriptor for a hand region may be used alone or in combination with each other and/or one of the example methods described above. For example, shape descriptors, such as the histograms or vectors described above, may be mixed and matched, combined, and/or concatenated into larger vectors, etc. This may allow the identification of new patterns that were not identifiable by looking at them in isolation. These filters may be augmented by the use of historical frame data, which can indicate whether an identified finger, for example, deviates too much from that finger identified in a previous frame.

[0092] Fig. 18 shows a supervisor filter for combining the results of various filters described above. For example, the pixel classifier filter may be used to produce a model of the hand and fingers. Moreover, the pixel classifier, the curvature analysis filter, the depth histogram filter, and possibly other hand filters not shown in Fig. 19 may be processed as described above, and further processed, for example by temporal consistency filtering (e.g., a low pass filter) and smoothing techniques, to produce hand and finger positions. As mentioned above, the silhouette used in various filters described herein may be scaled to be hand-size and sensor-distance invariant through knowledge of the user's distance from the camera and hand-size inferred from their analyzed skeleton.

[0093] In addition to open or closed hand states, the present technology may be used to identify specific finger orientations, such as for example pointing in a particular direction with one or more fingers. The technology may also be used to identify various hand positions oriented at various angles within x, y, z Cartesian space.

5 [0094] In embodiments, various post-classification filtering steps may be employed to increase accuracy of the hand and finger position estimations in step 216 (Fig. 6). For example, a temporal-consistency filtering, step may be applied to predicted hand and finger positions between consecutive depth image frames to smooth the predictions and reduce temporal jittering, e.g., due to spurious hand movements, sensor noise, or
10 occasional classification errors. That is, a plurality of hand and finger positions based on a plurality of depth images from the capture device or sensor may be estimated and temporal filtering of the plurality of estimates to estimate the hand and finger positions may be performed.

[0095] In step 220, the pipeline of Fig. 6 may output a response each frame based on
15 the estimated hand state. For example, a command may be output to a console of a computing system, such as console 12 of computing system 10. As another example, a response may be output to a display device, such as display device 16. In this way, estimated motions of the user, including estimated hand states, may be translated into commands to a console 12 of the system 10, so that the user may interact with the system
20 as described above. Further, the method and processes described above may be implemented to determine estimates of states of any part of a user's body, e.g., mouth, eyes, etc. For example, a posture of a body part of a user may be estimated using the methods described above.

[0096] The present technology enables a wide variety of interactions with a NUI
25 system such as for example shown in Figs. 1A- 1C. There are a wide range of natural interactions that are based on hand/finger movements, or combine both large body motions and fine grained hand control that are desirable to create new recognized gestures, more immersive experiences and compelling games. These uses and interactions include but are not limited to the following:

- 30
- Providing high-fidelity cursor positions - by recognizing and tracking with precision a user's pointing finger, the NUI system may accurately determine where a user is pointing on the screen with respect to positioning a cursor (Fig. 1B).

- Directional aiming of fingers - in general, the precise recognition and tracking of a user's finger or fingers may be used in any of a variety of ways to improve control over and interaction with a NUI system and a gaming or other application running on the NUI system. The recognition of various hand configurations may be used as recognized gestures such as for example, but not limited to, counting with fingers, thumbs-up, thumbs-down, the "a-okay" sign, the horns sign (index and little finger pointing up), the "hang loose" sign, the Star Trek® "live long and prosper" sign, a single upraised finger, and others. Each of these may be used to drive user interface interaction.
- Virtual Buttons (with haptic feedback) - accurate recognition and tracking of individual fingers allows applications to use a variety of virtual buttons, further enhancing the NUI experience.
- Thumb and finger control - taking the orientation and reliable detection of the thumb from fingers, the hand can act as a controller - orientation of the thumb drives controller orientation, pressing the thumb to the hand is recognized as a button press.
- Pinching to select - accurate recognition and tracking of individual fingers allows applications to use a pinching motion between the thumb and another finger to affect some control function or application metric.
- Single/Multiple finger directions - accurate recognition and tracking of individual fingers allows applications to use the relative positions of fingers as a control metric or to accomplish some other application metric.
- Writing, drawing sculpting - accurate recognition and tracking of individual fingers allows applications to interpret a user holding a pen or paintbrush, and how that pen or paintbrush moves upon movement of individual fingers. Recognition of such movements enables a user to form letters, cursive script, sculpt and/or draw images.
- Typing - accurate recognition and tracking of individual fingers allows applications to perform typing motions which are interpreted by the NUI system or application as key strokes on a virtual keyboard to type character and words on the screen or to provide control or application information to the NUI system or application.

- Tracking hand rotations - accurate recognition and tracking of individual fingers allows applications to accurately identify hand rotation.
- Puppeteering - mapping the finger skeleton to a puppeteer animation control system. Alternatively, mapping a finger skeleton may be used to directly
5 manipulate a virtual object in the same form as a physical puppet on a physical string is manipulated
- Turning a knob or combination lock - accurate recognition and tracking of individual fingers allows a user to select and twist a virtual knob or open a virtual combination lock. Such a combination lock may be used to provide or deny access
10 to secure network or stored resources.
- Shooting a gun - using finger and hand detection as a gun controller – index finger determining aim and thumb-press as a button to indicate shooting.
- Flicking Gesture - detecting and using an in-air finger flicking gesture for virtual interaction.
- Open palm gesture - Using an open palm to display a map view signifying a modal
15 change between first person and third person point of view. An index finger can be used on the open palm (similar to a mouse, or a finger on a touch screen) to scroll across and navigate through the virtual space.
- Leg control - Using the index and middle finger (with the hand pointing down) to
20 control the legs of a character, simulating running, jumping and kicking actions. This gesture could be combined with the open palm gesture to signify a modal change between full body interactions and user interface, or navigation. For example, in an action adventure game a player could use full body controls to engage in combat, then use an open palm gesture to switch to map view and use
25 index and middle finger simulating running for terrain traversal.

[0097] Other finger and hand interactions are contemplated.

[0098] Fig. 19A illustrates an example embodiment of a computing environment that may be used to interpret one or more positions and motions of a user in a target recognition, analysis, and tracking system. The computing environment such as the
30 computing environment 12 described above with respect to Figs. 1A-2 may be a multimedia console 600, such as a gaming console. As shown in Fig. 19A, the multimedia console 600 has a central processing unit (CPU) 601 having a level 1 cache 602, a level 2 cache 604, and a flash ROM 606. The level 1 cache 602 and a level 2 cache 604

temporarily store data and hence reduce the number of memory access cycles, thereby improving processing speed and throughput. The CPU 601 may be provided having more than one core, and thus, additional level 1 and level 2 caches 602 and 604. The flash ROM 606 may store executable code that is loaded during an initial phase of a boot process when the multimedia console 600 is powered ON.

5 [0099] A graphics processing unit (GPU) 608 and a video encoder/video codec (coder/decoder) 614 form a video processing pipeline for high speed and high resolution graphics processing. Data is carried from the GPU 608 to the video encoder/video codec 614 via a bus. The video processing pipeline outputs data to an A/V (audio/video) port 10 640 for transmission to a television or other display. A memory controller 610 is connected to the GPU 608 to facilitate processor access to various types of memory 612, such as, but not limited to, a RAM.

[00100] The multimedia console 600 includes an I/O controller 620, a system management controller 622, an audio processing unit 623, a network interface controller 15 624, a first USB host controller 626, a second USB host controller 628 and a front panel I/O subassembly 630 that are preferably implemented on a module 618. The USB controllers 626 and 628 serve as hosts for peripheral controllers 642(1)-642(2), a wireless adapter 648, and an external memory device 646 (e.g., flash memory, external CD/DVD ROM drive, removable media, etc.). The network interface 624 and/or wireless adapter 20 648 provide access to a network (e.g., the Internet, home network, etc.) and may be any of a wide variety of various wired or wireless adapter components including an Ethernet card, a modem, a Bluetooth module, a cable modem, and the like.

[00101] System memory 643 is provided to store application data that is loaded during the boot process. A media drive 644 is provided and may comprise a DVD/CD drive, hard 25 drive, or other removable media drive, etc. The media drive 644 may be internal or external to the multimedia console 600. Application data may be accessed via the media drive 644 for execution, playback, etc. by the multimedia console 600. The media drive 644 is connected to the I/O controller 620 via a bus, such as a Serial ATA bus or other high speed connection (e.g., IEEE 1394).

30 [00102] The system management controller 622 provides a variety of service functions related to assuring availability of the multimedia console 600. The audio processing unit 623 and an audio codec 632 form a corresponding audio processing pipeline with high fidelity and stereo processing. Audio data is carried between the audio processing unit 623 and the audio codec 632 via a communication link. The audio processing pipeline

outputs data to the A/V port 640 for reproduction by an external audio player or device having audio capabilities.

5 [00103] The front panel I/O subassembly 630 supports the functionality of the power button 650 and the eject button 652, as well as any LEDs (light emitting diodes) or other indicators exposed on the outer surface of the multimedia console 600. A system power supply module 636 provides power to the components of the multimedia console 600. A fan 638 cools the circuitry within the multimedia console 600.

10 [00104] The CPU 601, GPU 608, memory controller 610, and various other components within the multimedia console 600 are interconnected via one or more buses, including serial and parallel buses, a memory bus, a peripheral bus, and a processor or local bus using any of a variety of bus architectures. By way of example, such architectures can include a Peripheral Component Interconnects (PCI) bus, PCI-Express bus, etc.

15 [00105] When the multimedia console 600 is powered ON, application data may be loaded from the system memory 643 into memory 612 and/or caches 602, 604 and executed on the CPU 601. The application may present a graphical user interface that provides a consistent user experience when navigating to different media types available on the multimedia console 600. In operation, applications and/or other media contained within the media drive 644 may be launched or played from the media drive 644 to provide additional functionalities to the multimedia console 600.

20 [00106] The multimedia console 600 may be operated as a standalone system by simply connecting the system to a television or other display. In this standalone mode, the multimedia console 600 allows one or more users to interact with the system, watch movies, or listen to music. However, with the integration of broadband connectivity made available through the network interface 624 or the wireless adapter 648, the multimedia console 600 may further be operated as a participant in a larger network community.

25 [00107] When the multimedia console 600 is powered ON, a set amount of hardware resources are reserved for system use by the multimedia console operating system. These resources may include a reservation of memory (e.g., 16MB), CPU and GPU cycles (e.g., 5%), networking bandwidth (e.g., 8 kbs), etc. Because these resources are reserved at system boot time, the reserved resources do not exist from the application's view.

30 [00108] In particular, the memory reservation preferably is large enough to contain the launch kernel, concurrent system applications and drivers. The CPU reservation is preferably constant such that if the reserved CPU usage is not used by the system

applications, an idle thread will consume any unused cycles.

[00109] With regard to the GPU reservation, lightweight messages generated by the system applications (e.g., popups) are displayed by using a GPU interrupt to schedule code to render popup into an overlay. The amount of memory required for an overlay depends
5 on the overlay area size and the overlay preferably scales with screen resolution. Where a full user interface is used by the concurrent system application, it is preferable to use a resolution independent of the application resolution. A scaler may be used to set this resolution such that the need to change frequency and cause a TV resynch is eliminated.

[00110] After the multimedia console 600 boots and system resources are reserved,
10 concurrent system applications execute to provide system functionalities. The system functionalities are encapsulated in a set of system applications that execute within the reserved system resources described above. The operating system kernel identifies threads that are system application threads versus gaming application threads. The system applications are preferably scheduled to run on the CPU 601 at predetermined times and
15 intervals in order to provide a consistent system resource view to the application. The scheduling is to minimize cache disruption for the gaming application running on the console.

[00111] When a concurrent system application requires audio, audio processing is scheduled asynchronously to the gaming application due to time sensitivity. A multimedia
20 console application manager (described below) controls the gaming application audio level (e.g., mute, attenuate) when system applications are active.

[00112] Input devices (e.g., controllers 642(1) and 642(2)) are shared by gaming applications and system applications. The input devices are not reserved resources, but are to be switched between system applications and the gaming application such that each will
25 have a focus of the device. The application manager preferably controls the switching of input stream, without knowledge of the gaming application's knowledge and a driver maintains state information regarding focus switches. The cameras 26, 28 and capture device 20 may define additional input devices for the console 600.

[00113] Fig. 19B illustrates another example embodiment of a computing environment
30 720 that may be the computing environment 12 shown in Figs. 1A-2 used to interpret one or more positions and motions in a target recognition, analysis, and tracking system. The computing system environment 720 is only one example of a suitable computing environment and is not intended to suggest any limitation as to the scope of use or functionality of the presently disclosed subject matter. Neither should the computing

environment 720 be interpreted as having any dependency or requirement relating to any one or combination of components illustrated in the Exemplary operating environment 720. In some embodiments, the various depicted computing elements may include circuitry configured to instantiate specific aspects of the present disclosure. For example, the term circuitry used in the disclosure can include specialized hardware components configured to perform function(s) by firmware or switches. In other example embodiments, the term circuitry can include a general purpose processing unit, memory, etc., configured by software instructions that embody logic operable to perform function(s). In example embodiments where circuitry includes a combination of hardware and software, an implementer may write source code embodying logic and the source code can be compiled into machine readable code that can be processed by the general purpose processing unit. Since one skilled in the art can appreciate that the state of the art has evolved to a point where there is little difference between hardware, software, or a combination of hardware/software, the selection of hardware versus software to effectuate specific functions is a design choice left to an implementer. More specifically, one of skill in the art can appreciate that a software process can be transformed into an equivalent hardware structure, and a hardware structure can itself be transformed into an equivalent software process. Thus, the selection of a hardware implementation versus a software implementation is one of design choice and left to the implementer.

[00114] In Fig. 19B, the computing environment 720 comprises a computer 741, which typically includes a variety of computer readable media. Computer readable media can be any available media that can be accessed by computer 741 and includes both volatile and nonvolatile media, removable and non-removable media. The system memory 722 includes computer storage media in the form of volatile and/or nonvolatile memory such as ROM 723 and RAM 760. A basic input/output system 724 (BIOS), containing the basic routines that help to transfer information between elements within computer 741, such as during start-up, is typically stored in ROM 723. RAM 760 typically contains data and/or program modules that are immediately accessible to and/or presently being operated on by processing unit 759. By way of example, and not limitation, Fig. 19B illustrates operating system 725, application programs 726, other program modules 727, and program data 728. Fig. 19B further includes a graphics processor unit (GPU) 729 having an associated video memory 730 for high speed and high resolution graphics processing and storage. The GPU 729 may be connected to the system bus 721 through a graphics interface 731.

[00115] The computer 741 may also include other removable/non-removable, volatile/nonvolatile computer storage media. By way of example only, Fig. 19B illustrates a hard disk drive 738 that reads from or writes to non-removable, nonvolatile magnetic media, a magnetic disk drive 739 that reads from or writes to a removable, nonvolatile magnetic disk 754, and an optical disk drive 740 that reads from or writes to a removable, nonvolatile optical disk 753 such as a CD ROM or other optical media. Other removable/non-removable, volatile/nonvolatile computer storage media that can be used in the Exemplary operating environment include, but are not limited to, magnetic tape cassettes, flash memory cards, digital versatile disks, digital video tape, solid state RAM, solid state ROM, and the like. The hard disk drive 738 is typically connected to the system bus 721 through a non-removable memory interface such as interface 734, and magnetic disk drive 739 and optical disk drive 740 are typically connected to the system bus 721 by a removable memory interface, such as interface 735.

[00116] The drives and their associated computer storage media discussed above and illustrated in Fig. 19B, provide storage of computer readable instructions, data structures, program modules and other data for the computer 741. In Fig. 19B, for example, hard disk drive 738 is illustrated as storing operating system 758, application programs 757, other program modules 756, and program data 755. Note that these components can either be the same as or different from operating system 725, application programs 726, other program modules 727, and program data 728. Operating system 758, application programs 757, other program modules 756, and program data 755 are given different numbers here to illustrate that, at a minimum, they are different copies. A user may enter commands and information into the computer 741 through input devices such as a keyboard 751 and a pointing device 752, commonly referred to as a mouse, trackball or touch pad. Other input devices (not shown) may include a microphone, joystick, game pad, satellite dish, scanner, or the like. These and other input devices are often connected to the processing unit 759 through a user input interface 736 that is coupled to the system bus, but may be connected by other interface and bus structures, such as a parallel port, game port or a universal serial bus (USB). The cameras 26, 28 and capture device 20 may define additional input devices for the console 700. A monitor 742 or other type of display device is also connected to the system bus 721 via an interface, such as a video interface 732. In addition to the monitor, computers may also include other peripheral output devices such as speakers 744 and printer 743, which may be connected through an output peripheral interface 733.

[00117] The computer 741 may operate in a networked environment using logical connections to one or more remote computers, such as a remote computer 746. The remote computer 746 may be a personal computer, a server, a router, a network PC, a peer device or other common network node, and typically includes many or all of the elements
5 described above relative to the computer 741, although only a memory storage device 747 has been illustrated in Fig. 19B. The logical connections depicted in Fig. 19B include a local area network (LAN) 745 and a wide area network (WAN) 749, but may also include other networks. Such networking environments are commonplace in offices, enterprise-wide computer networks, intranets and the Internet.

10 [00118] When used in a LAN networking environment, the computer 741 is connected to the LAN 745 through a network interface or adapter 737. When used in a WAN networking environment, the computer 741 typically includes a modem 750 or other means for establishing communications over the WAN 749, such as the Internet. The modem 750, which may be internal or external, may be connected to the system bus 721
15 via the user input interface 736, or other appropriate mechanism. In a networked environment, program modules depicted relative to the computer 741, or portions thereof, may be stored in the remote memory storage device. By way of example, and not limitation, Fig. 19B illustrates remote application programs 748 as residing on memory device 747. It will be appreciated that the network connections shown are Exemplary and
20 other means of establishing a communications link between the computers may be used.

[00119] The foregoing detailed description of the inventive system has been presented for purposes of illustration and description. It is not intended to be exhaustive or to limit the inventive system to the precise form disclosed. Many modifications and variations are possible in light of the above teaching. The described embodiments were chosen in order
25 to best explain the principles of the inventive system and its practical application to thereby enable others skilled in the art to best utilize the inventive system in various embodiments and with various modifications as are suited to the particular use contemplated. It is intended that the scope of the inventive system be defined by the claims appended hereto.

30

CLAIMS:

1. A method for generating a model of a user's hand including one or more fingers, comprising:

5 (a) receiving position data representing a position of a user interacting with a sensor, the position data including at least one of depth and image data representing the user's hand; and

(b) analyzing the position data to identify the user's hand in the position data, said step (b) including the steps of:

10 (b)(1) analyzing depth data from the position data captured in said step (a) to segment the position data into data of the hand, and

(b)(2) extracting a set of feature descriptors by applying one or more filters to the image data of the hand identified in said step (b)(1), the one or more filters analyzing image data of the hand as compared to image data outside of a boundary of the hand to discern features of the hand including a shape and orientation of the hand.
15

2. The method of claim 1, further comprising the steps of running an application receiving commands via the sensing mechanism, and affecting a control action in the application based on a position of the hand identified in said step (b).

20 3. The method of claim 1, further comprising the steps of running a gaming application receiving commands via the sensing mechanism, and affecting an action in the gaming application based on a position of the hand identified in said step (b).

4. The method of claim 1, wherein said step (b)(1) comprises the step of analyzing centroids constructed from the image data to locate a best candidate of a hand.

25 5. The method of claim 4, wherein said step (b)(1) further comprises the step of analyzing the best candidate of a hand to determine a best candidate of the center of the hand.

6. The method of claim 1, wherein said step (b)(2) comprises the steps of applying a pixel classifier including the steps of:

- 5 selecting pixels within a boundary of the shape descriptor of the hand,
constructing a box of predetermined size around each pixel, each box constructed in a plane of the shape descriptor,
determining intersection points with each box where the image data transitions between a foreground point and a background point, and
identifying hand and fingers from analyzing the intersection points of each box for each examined pixel.

10 7. The method of claim 1, wherein said step (b)(2) comprises the steps of applying a curvature analysis filter including the steps of:

- 15 selecting pixels along a boundary of the shape descriptor of the hand,
examining a plurality of pixels surrounding a selected pixel, and assigning a value to the selected pixel indicating which surrounding pixel is also along a boundary of the shape descriptor,
converting the values into angles and contours around the hand including peaks and valleys, and
determining which of the peaks represent fingers of the hand.

20 8. The method of claim 1, wherein said step (b)(2) comprises the steps of applying a histogram filter including the step of constructing a histogram of distances between a plurality of points in the shape descriptor and a device capturing the image data.

9. A system for generating a model of a user's hand including one or more fingers, the system including a sensing mechanism operatively coupled to a computing device, the system comprising:

- a skeletal recognition engine for recognizing at least a portion of a skeleton of a user from received data including at least one of image and depth data;
- an image segmentation engine for segmenting one or more regions of the body into a region representing a hand of the user; and
- 5 a descriptor extraction engine for extracting data representative of a hand including one or more fingers and an orientation of the hand, the descriptor extraction engine applying a plurality of filters for analyzing pixels in the region representing the hand, each filter in the plurality of filters determining a position and orientation of the hand, the descriptor extraction engine combining the results of each filter to arrive at a best estimate of the
- 10 position and orientation of the hand.
10. The system of claim 9, wherein the plurality of filters of the descriptor extraction engine include one or more filters optimized for identifying the position and orientation of the hand as a silhouette relative to a device capturing the received data.
11. The system of claim 9, wherein the plurality of filters of the descriptor extraction engine
- 15 include one or more filters optimized for identifying the position and orientation of the hand when pointed toward or away from a device capturing the received data.
12. The system of claim 9, wherein the plurality of filters of the descriptor extraction engine include a classifier engine for analyzing a hand as a silhouette relative to the sensing mechanism, the classifier engine selecting pixels within a region representing a hand of the user, constructing a
- 20 box of predetermined size around each pixel, each box constructed in a plane of the silhouette, determining intersection points with each box where the image data transitions between a foreground point and a background point, and identifying hand and fingers from analyzing the intersection points of each box for each examined pixel.
13. The system of claim 12, wherein the classifier engine identifies a centroid representing a
- 25 fingertip where two intersection points are identified on a box and a distance between the intersection points is too small to represent a palm of the hand.
14. The system of claim 13, wherein a location of the two intersection points on the same or different sides of a box indicates an orientation of the identified fingertip.

15. The system of claim 12, wherein the classifier engine identifies a centroid representing a finger where four intersection points are identified on a box.

16. The system of claim 12, wherein the classifier engine identifies a centroid representing a palm of the hand where two intersection points are identified on a box and a distance between the
5 intersection points is too large to represent a fingertip.

17. The system of claim 12, wherein the box constructed around a given pixel is a first box of a first size, the pixel classifier engine further constructing a second box around the given pixel of a second size larger than the first size to detect a condition where fingers of the hand are together with each other.

10 18. A computer-readable medium having stored thereon computer-executable instructions, that when executed, perform a method according to any one of claims 1 to 8.

1/19

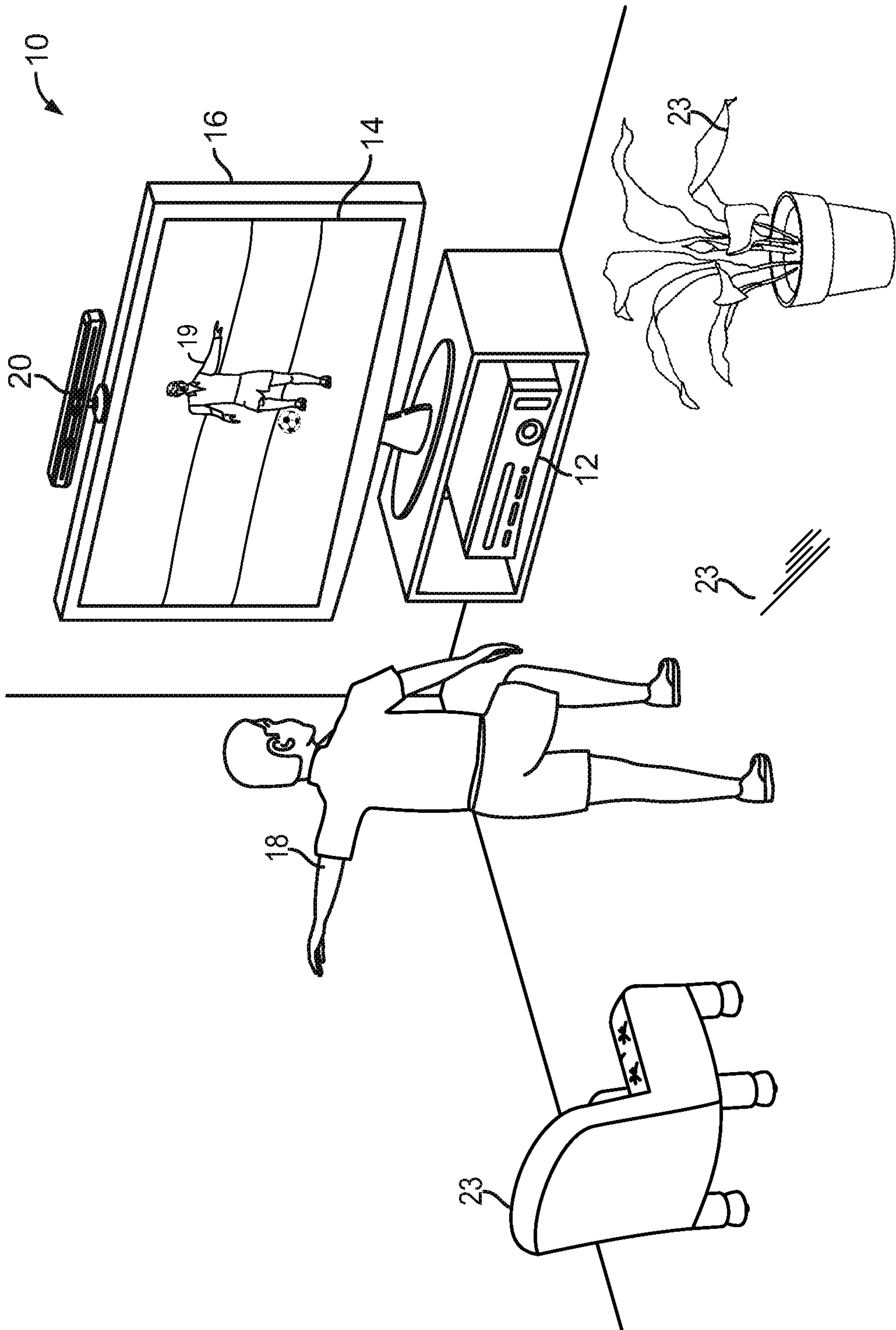


Fig. 1A

2/19

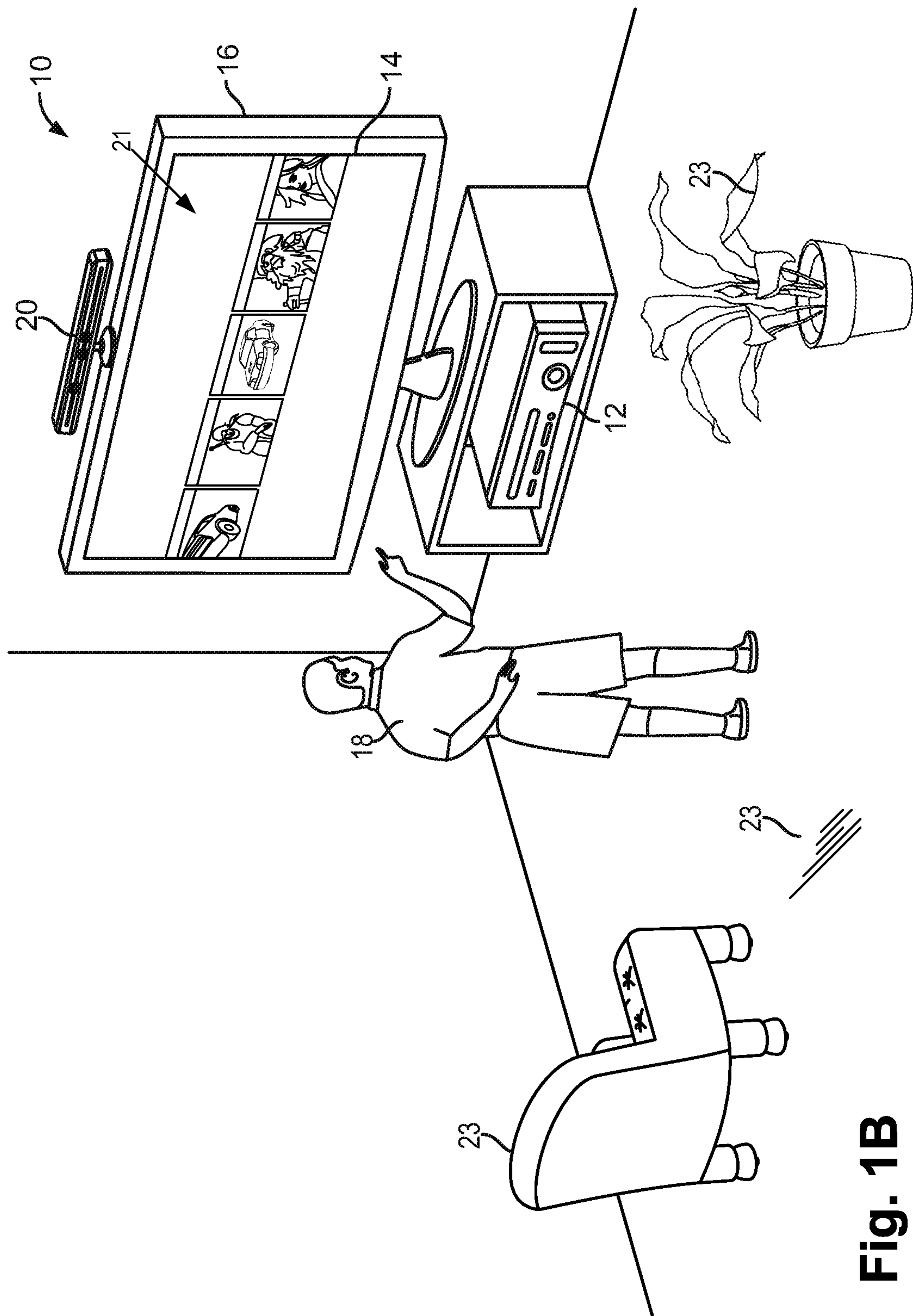


Fig. 1B

3/19

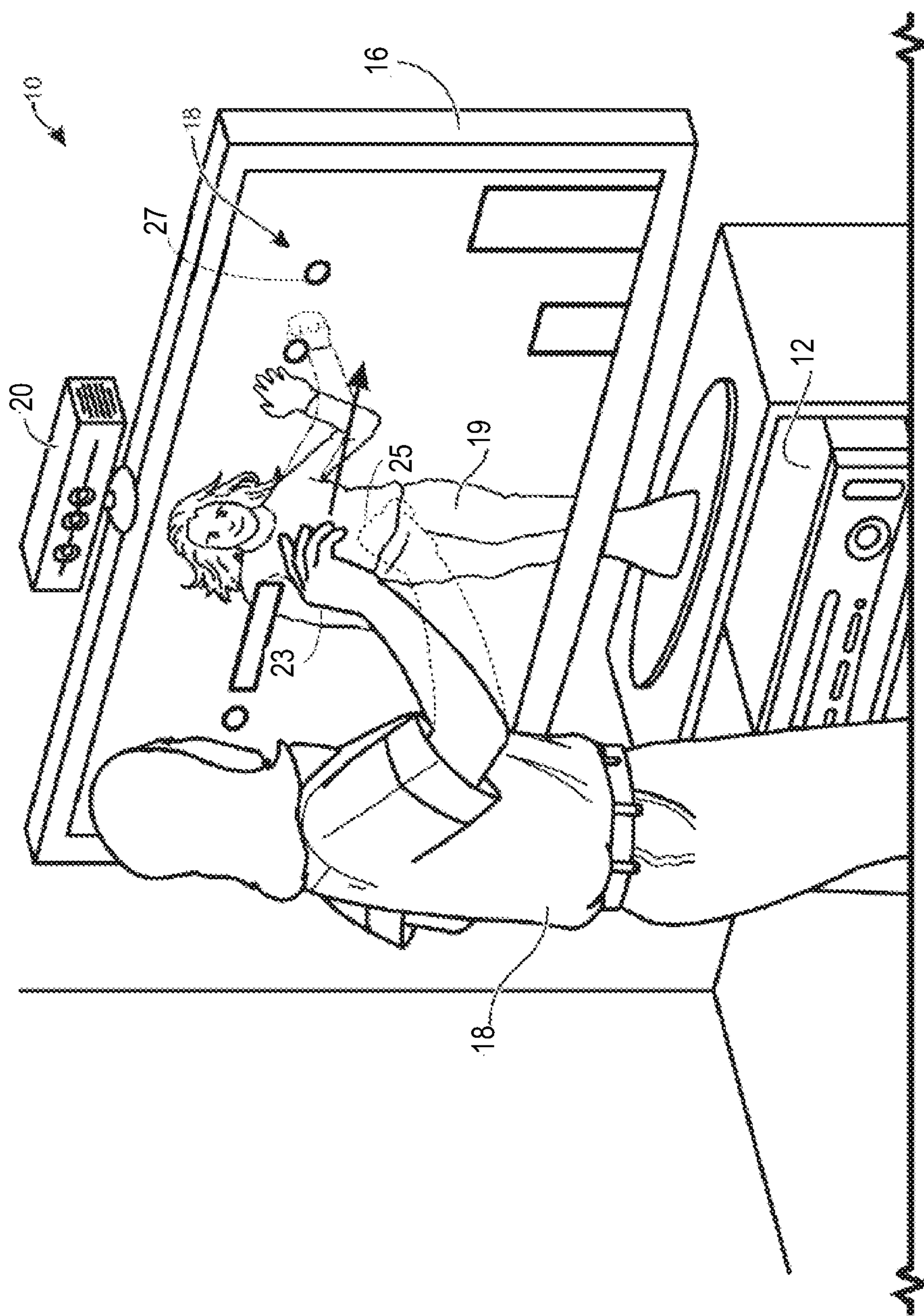


Fig. 1C

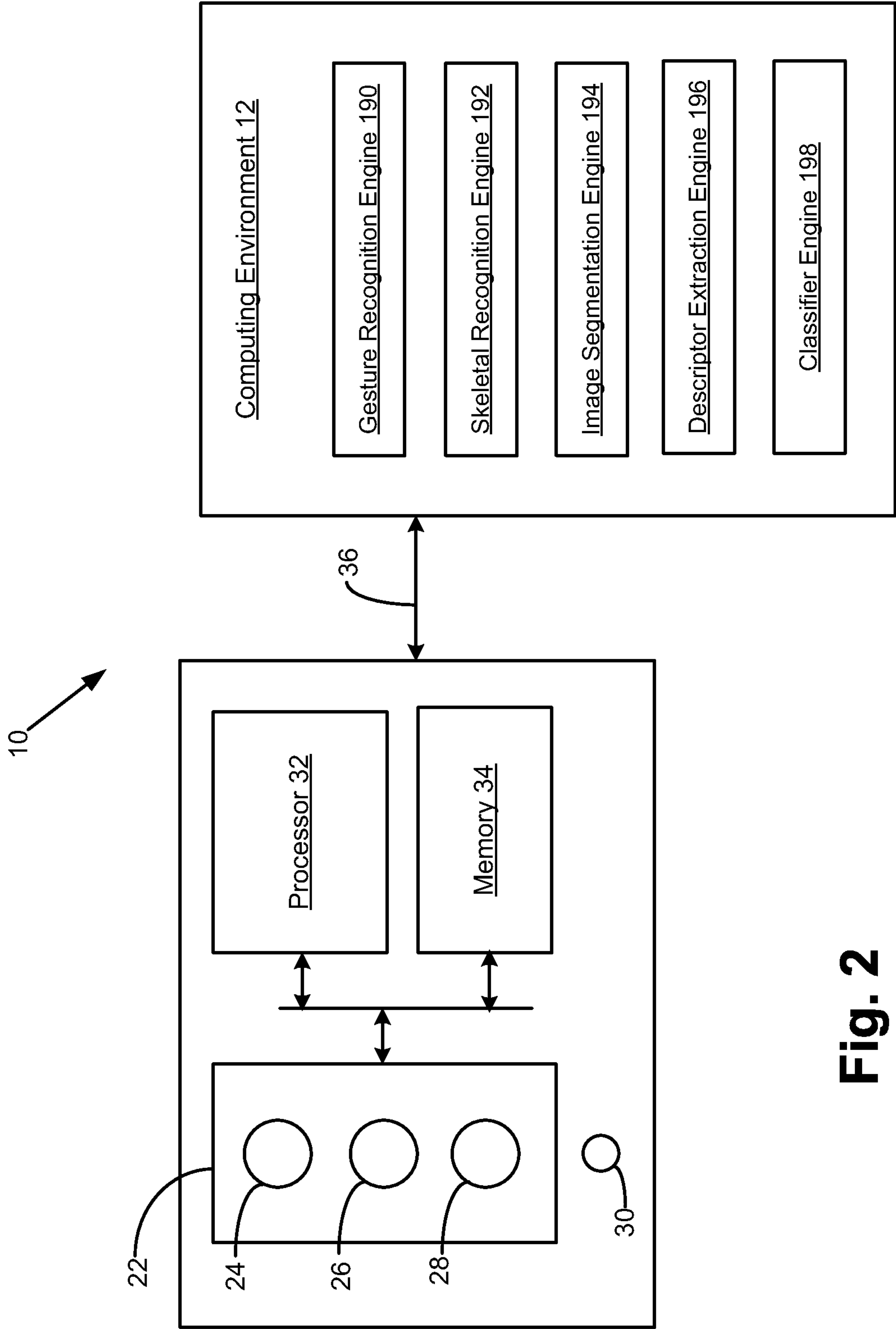
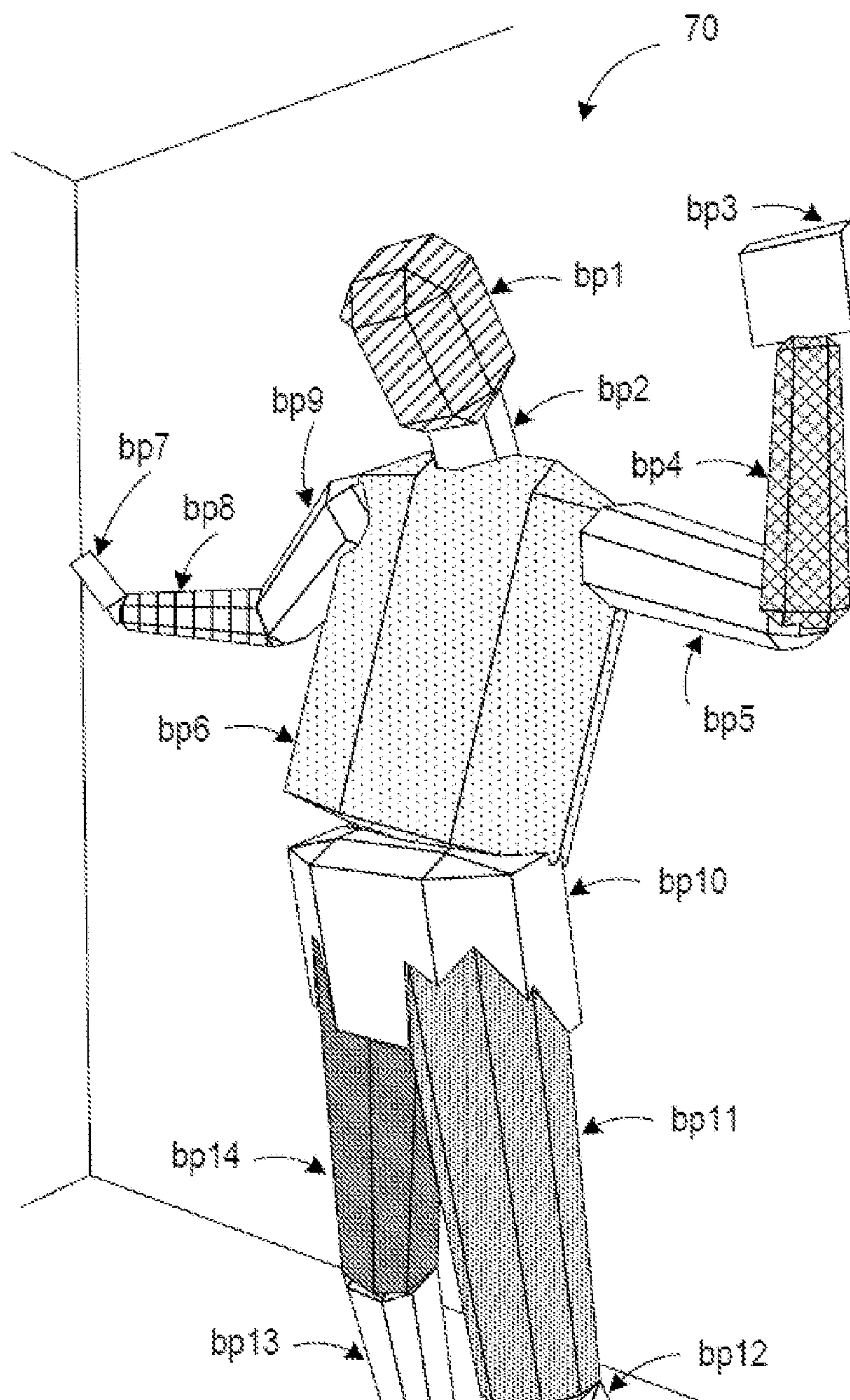


Fig. 2

5/19

**Fig. 3**

6/19

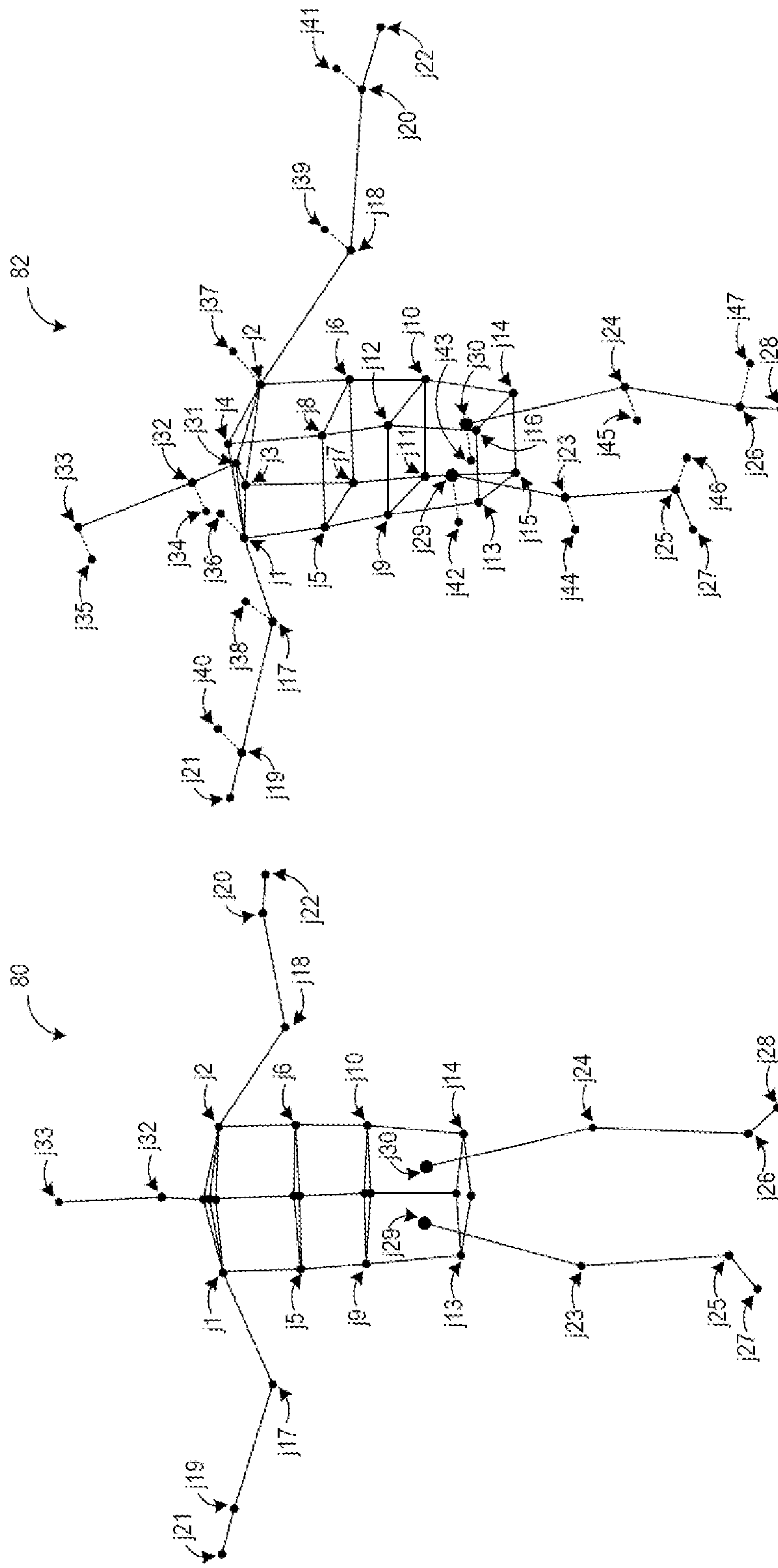
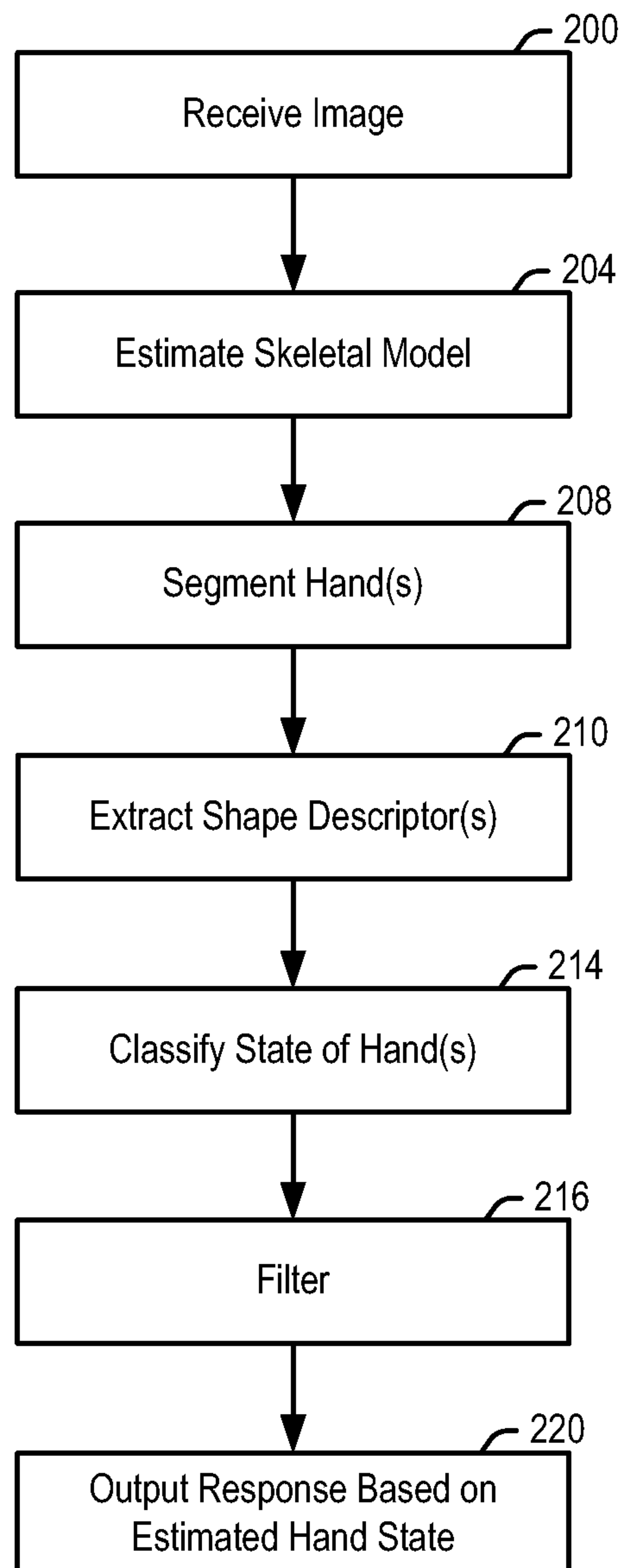


Fig. 4

Fig. 5

7/19

**Fig. 6**

8/19

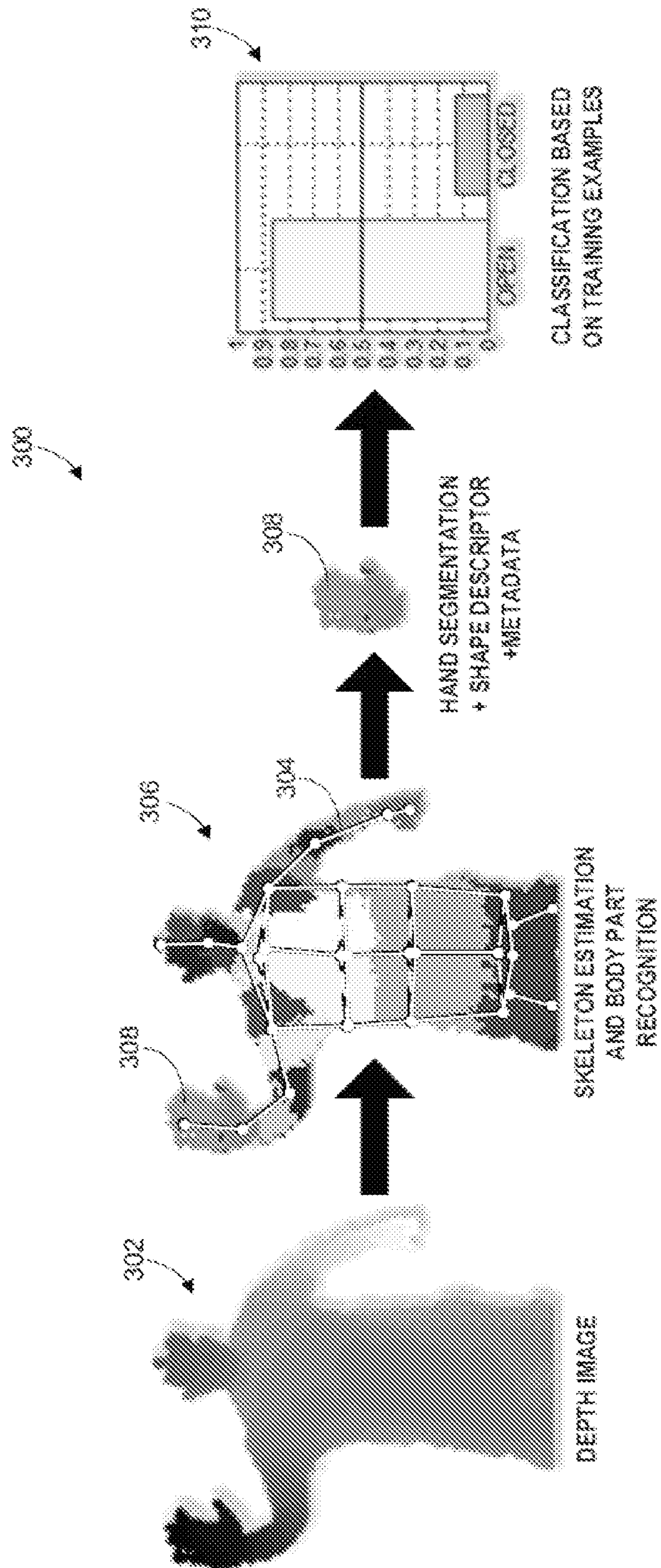
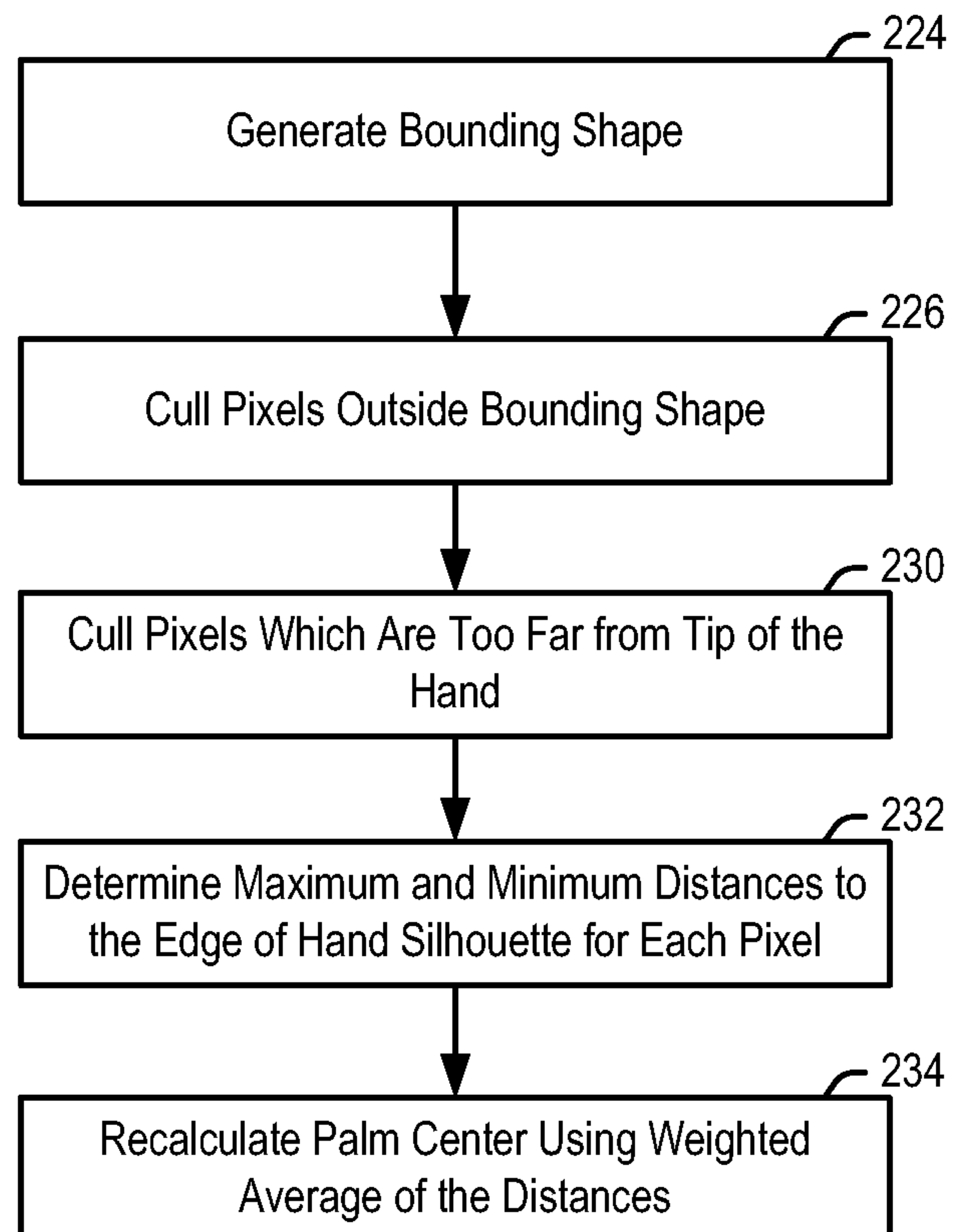
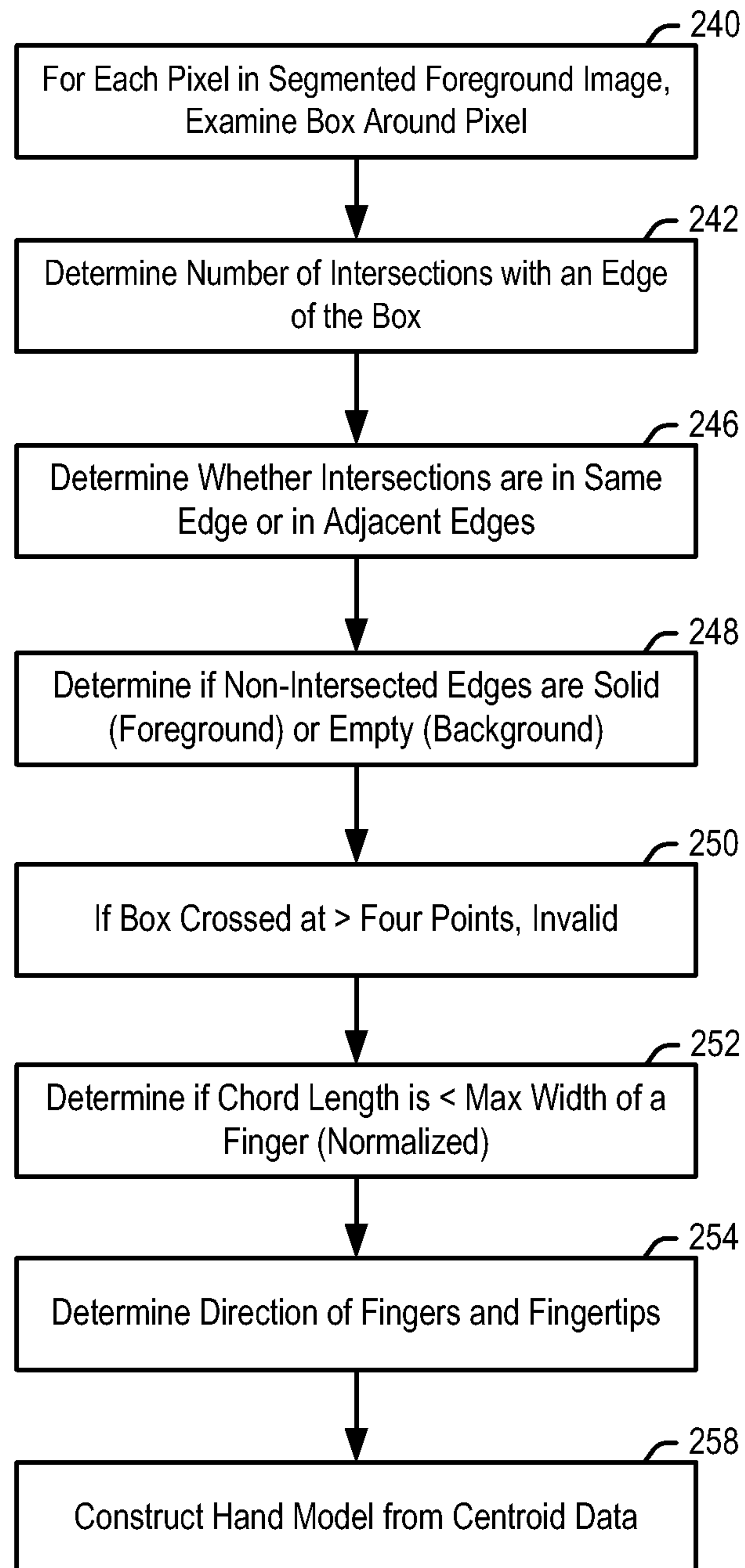


Fig. 7

9/19

**Fig. 8**

10/19**Fig. 9**

11/19

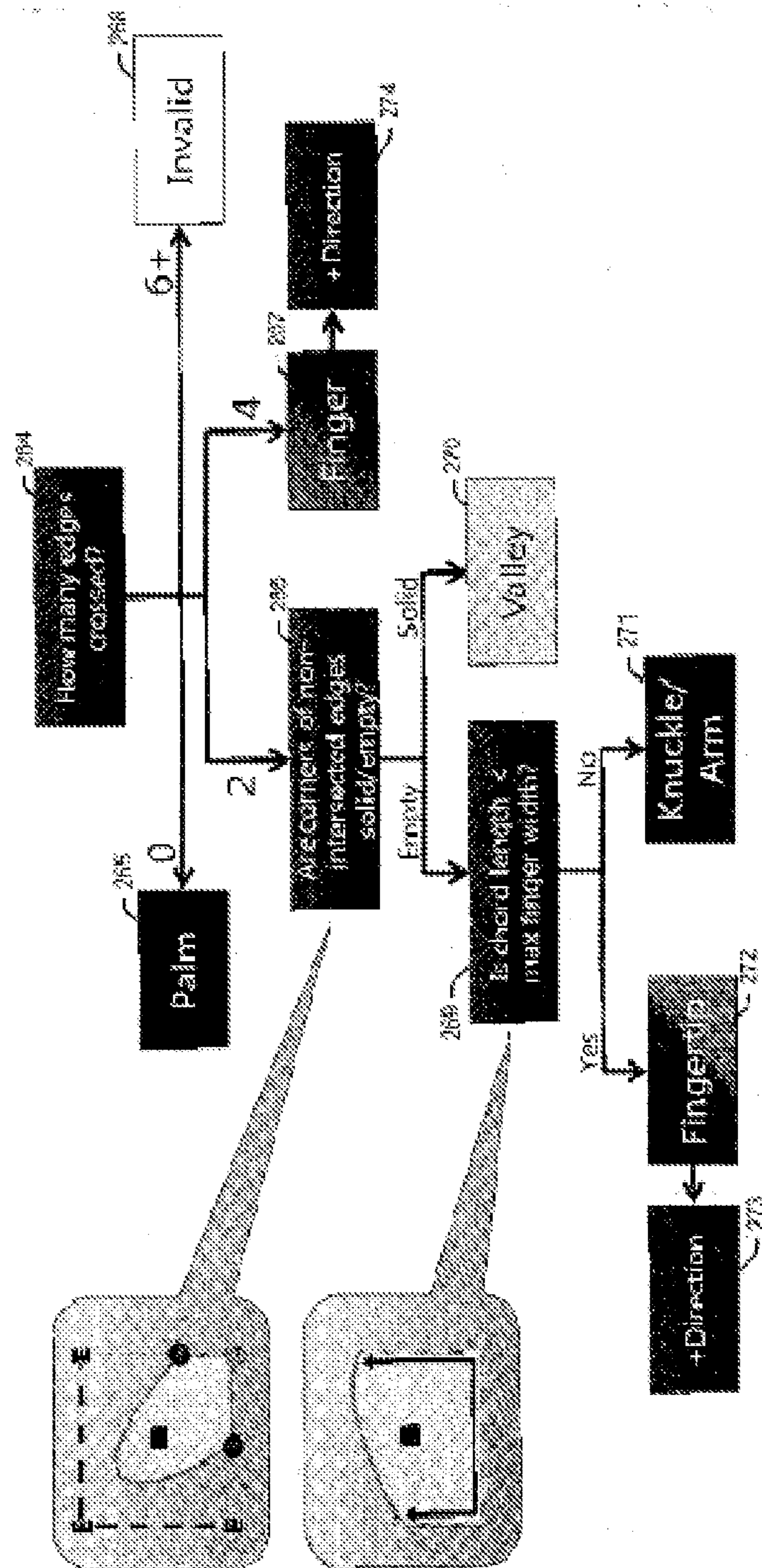
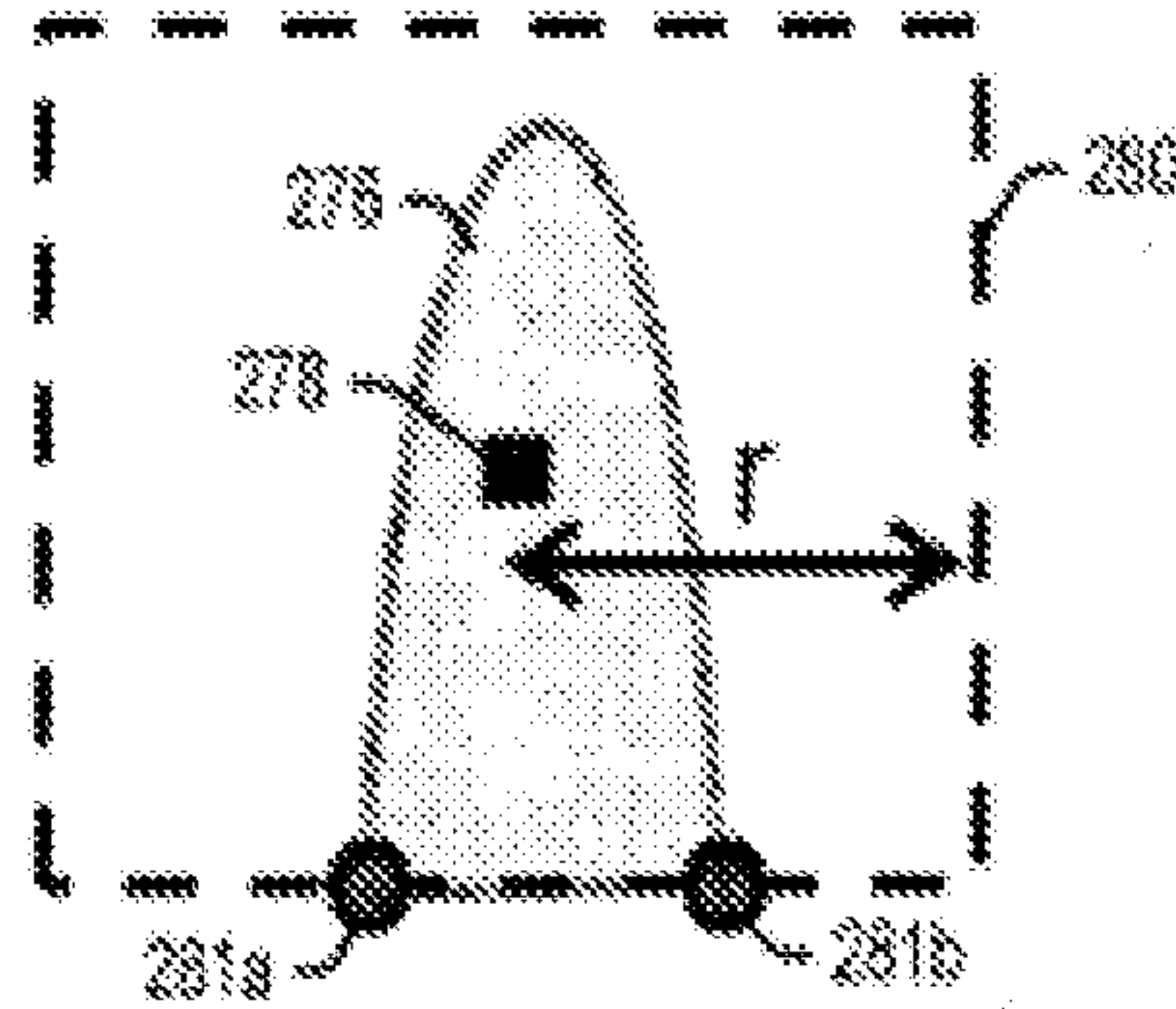
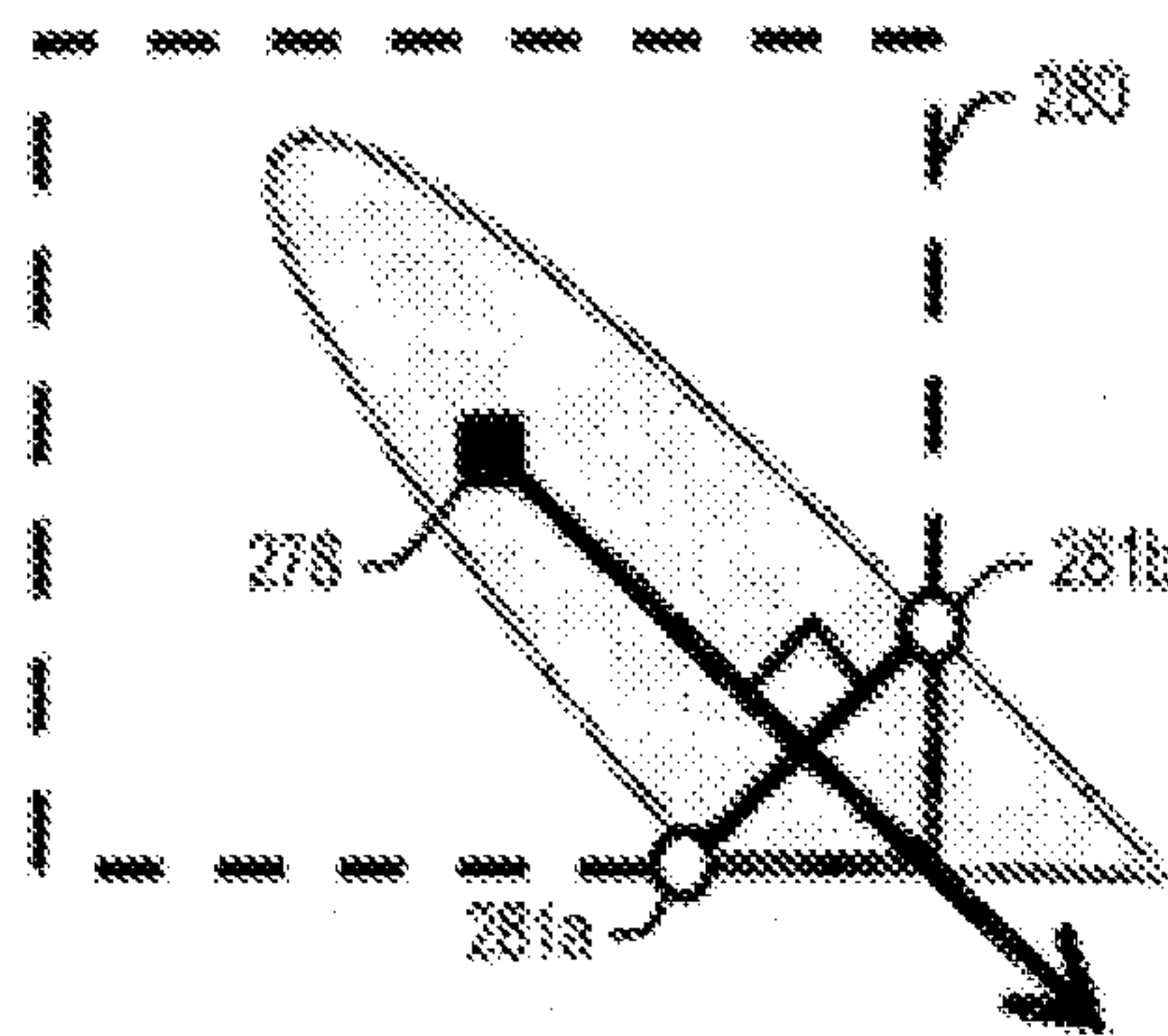
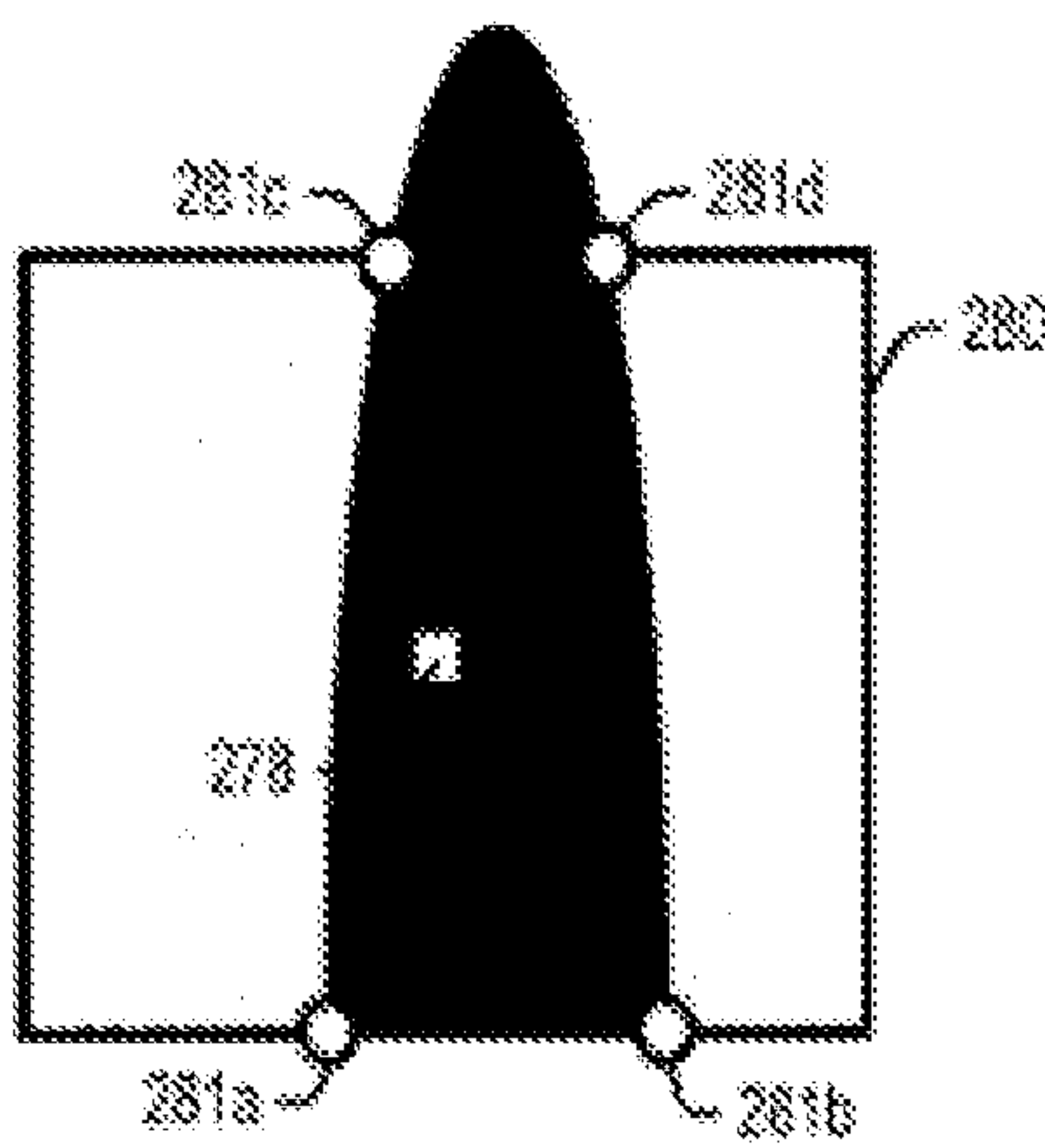
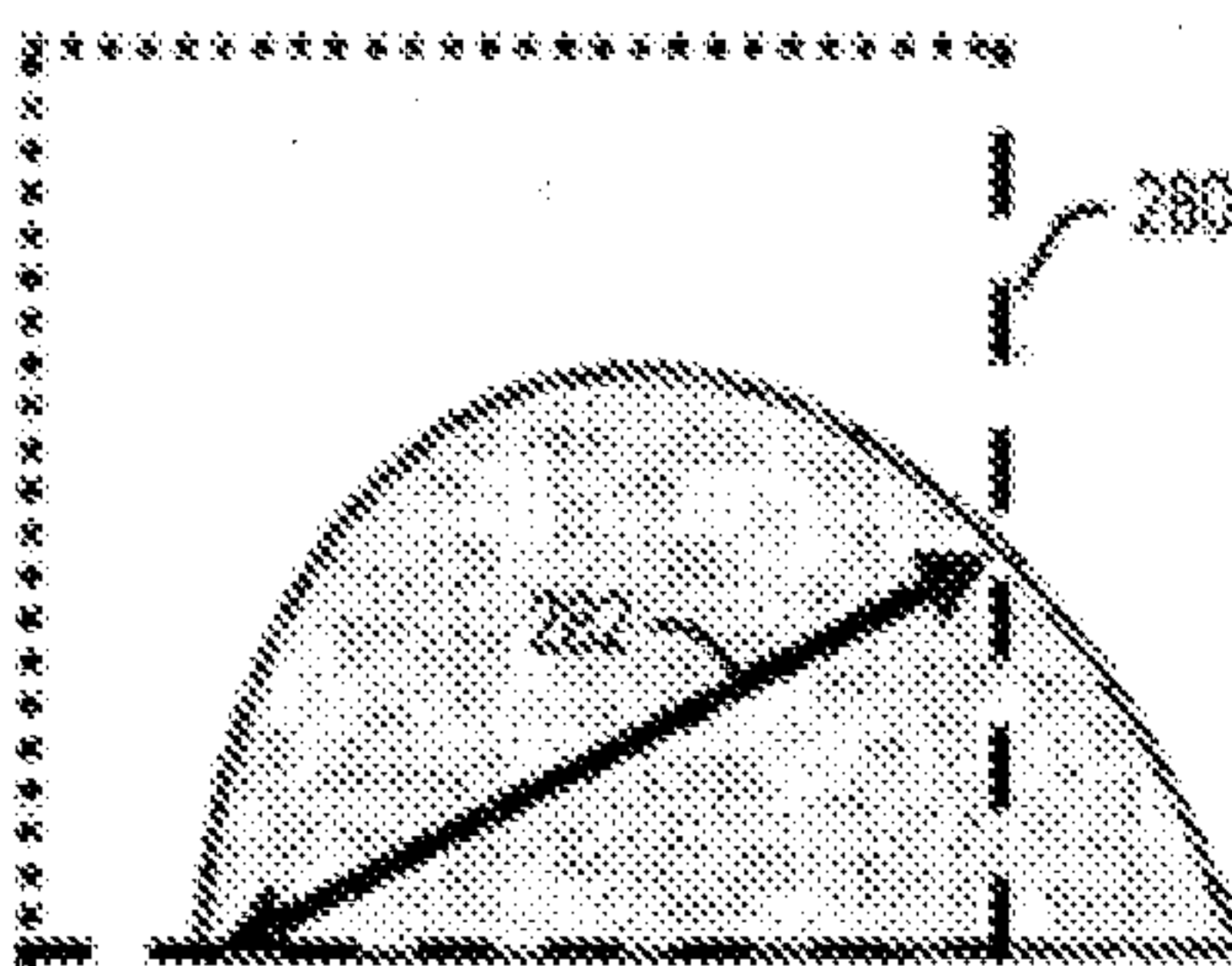
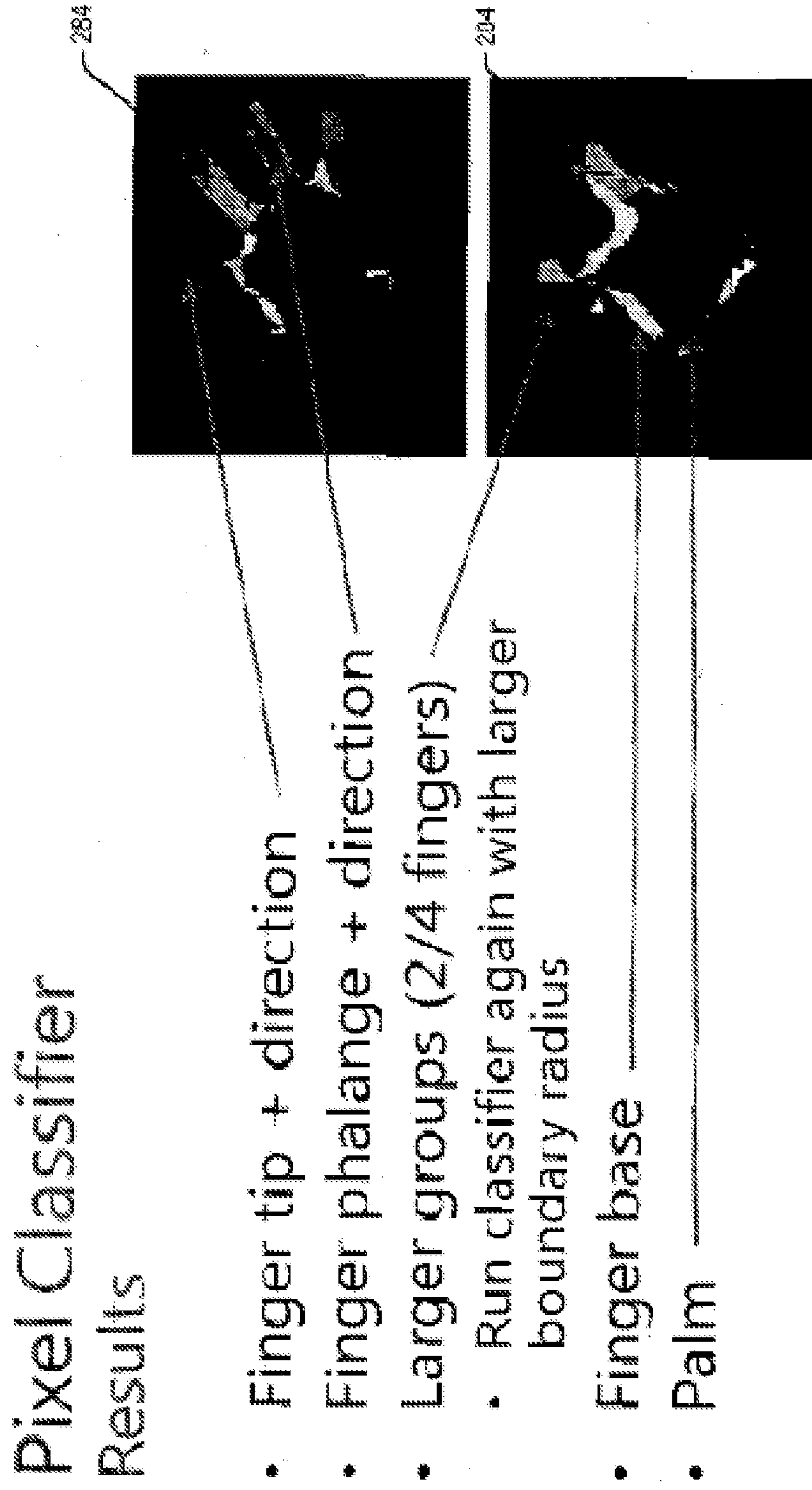


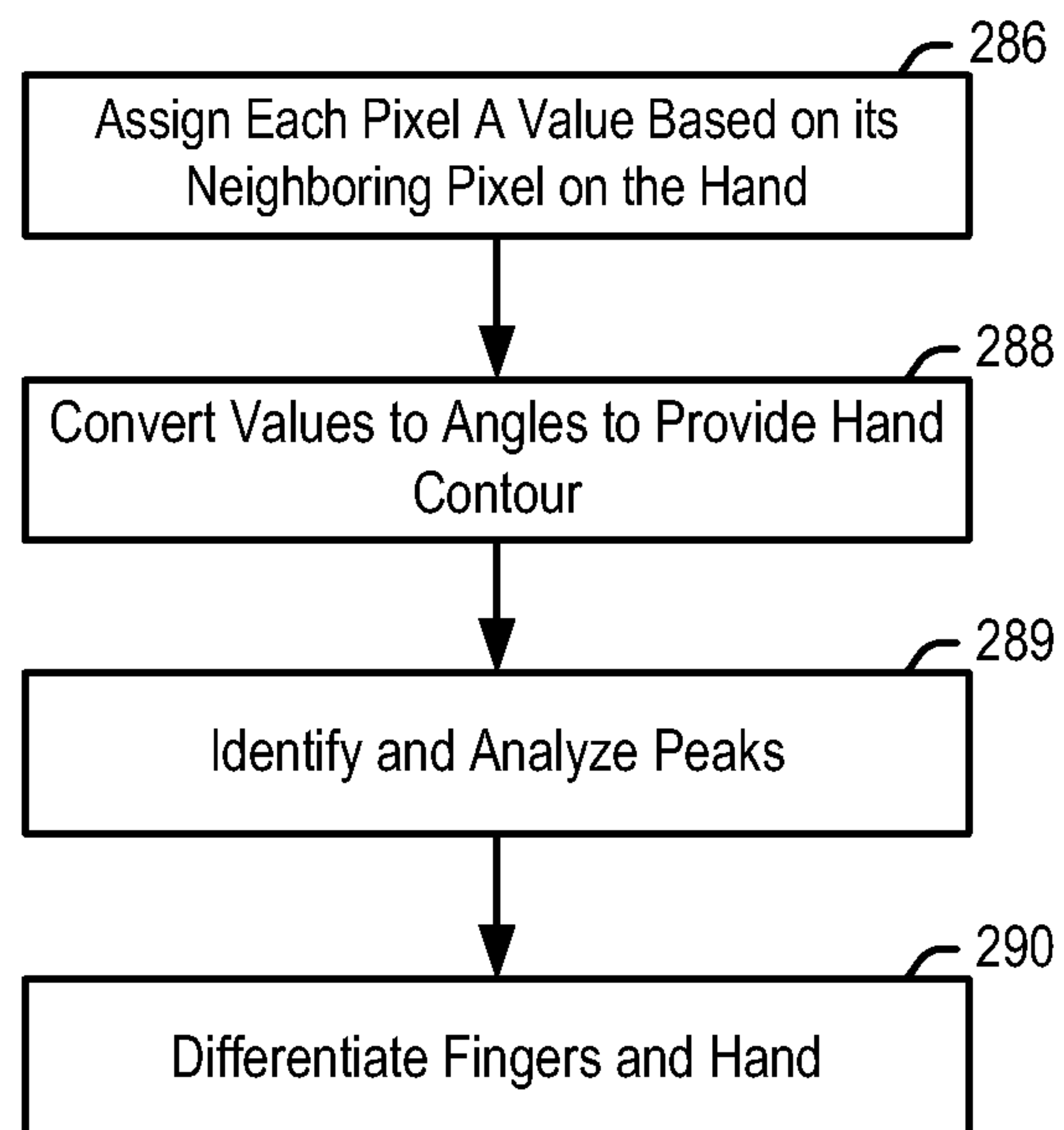
Fig. 10

12/19

Fig. 11A**Fig. 11B****Fig. 12****Fig. 13**

13/19

**Fig. 14**

14/19**Fig. 15**

15/19

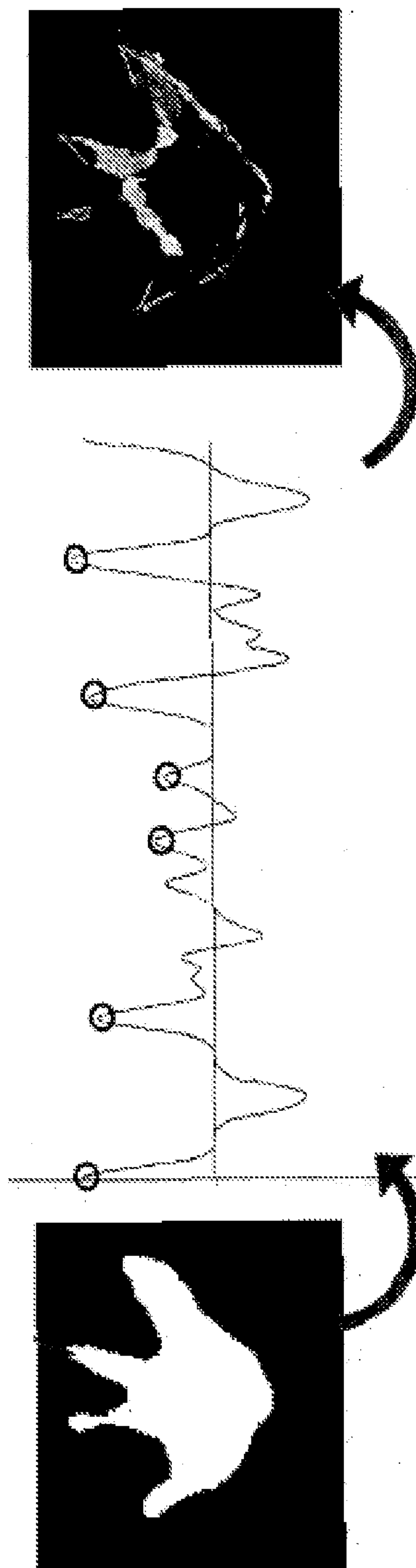
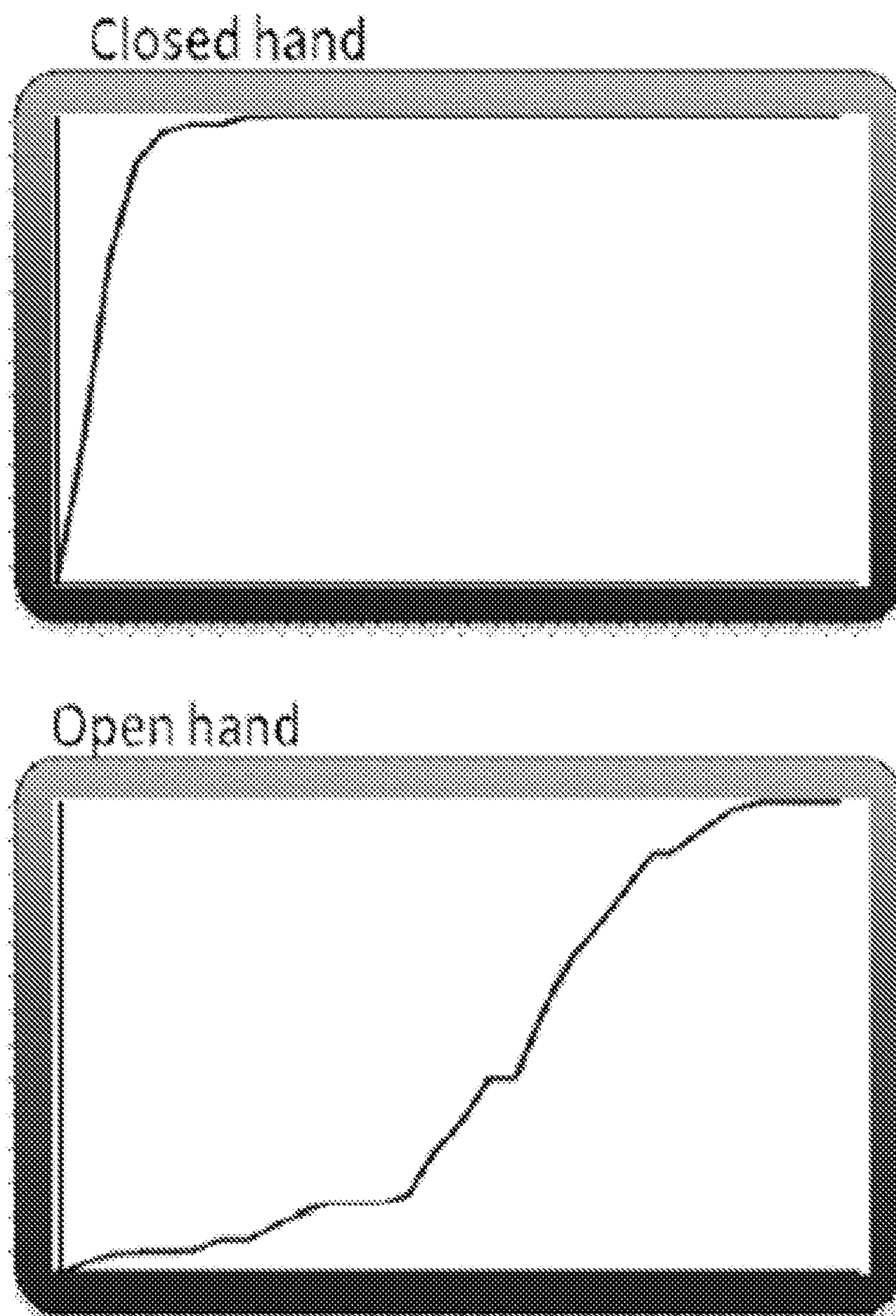


Fig. 16

16/19

**Fig. 17**

Supervisor
Combining Results

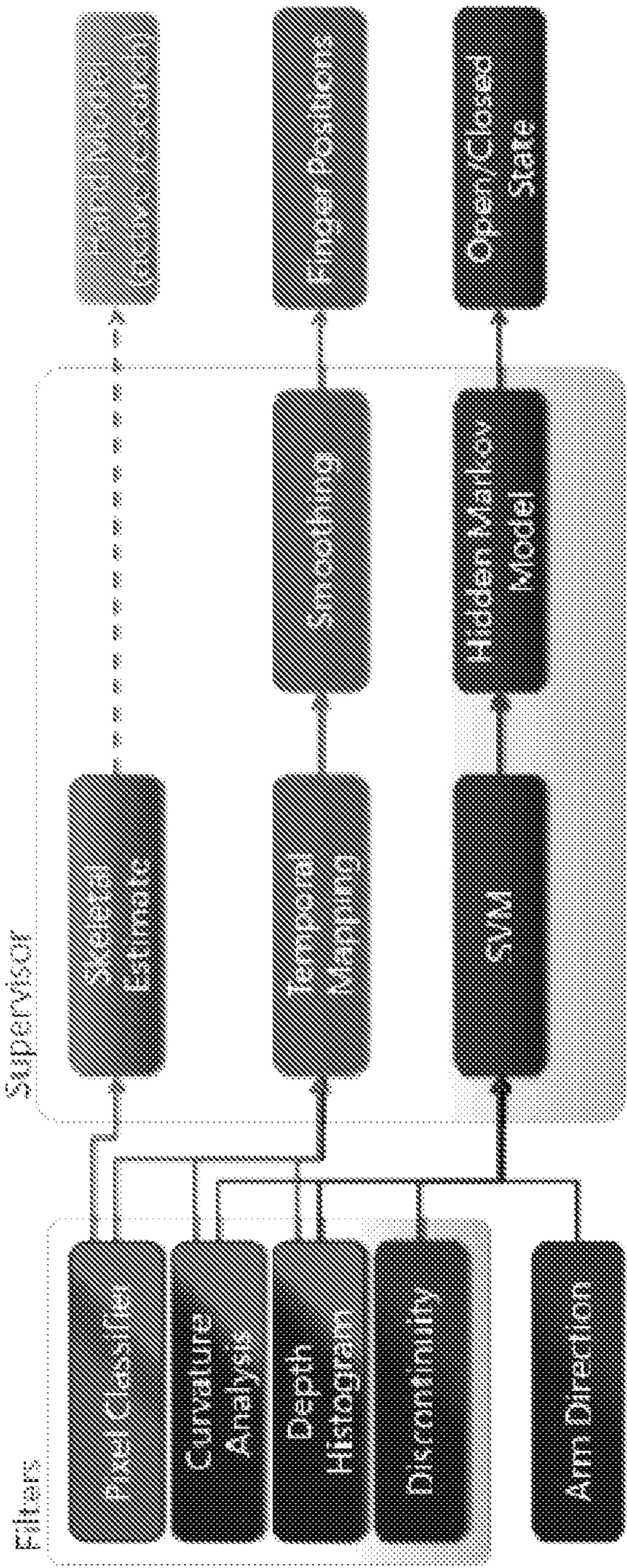


Fig. 18

18/19

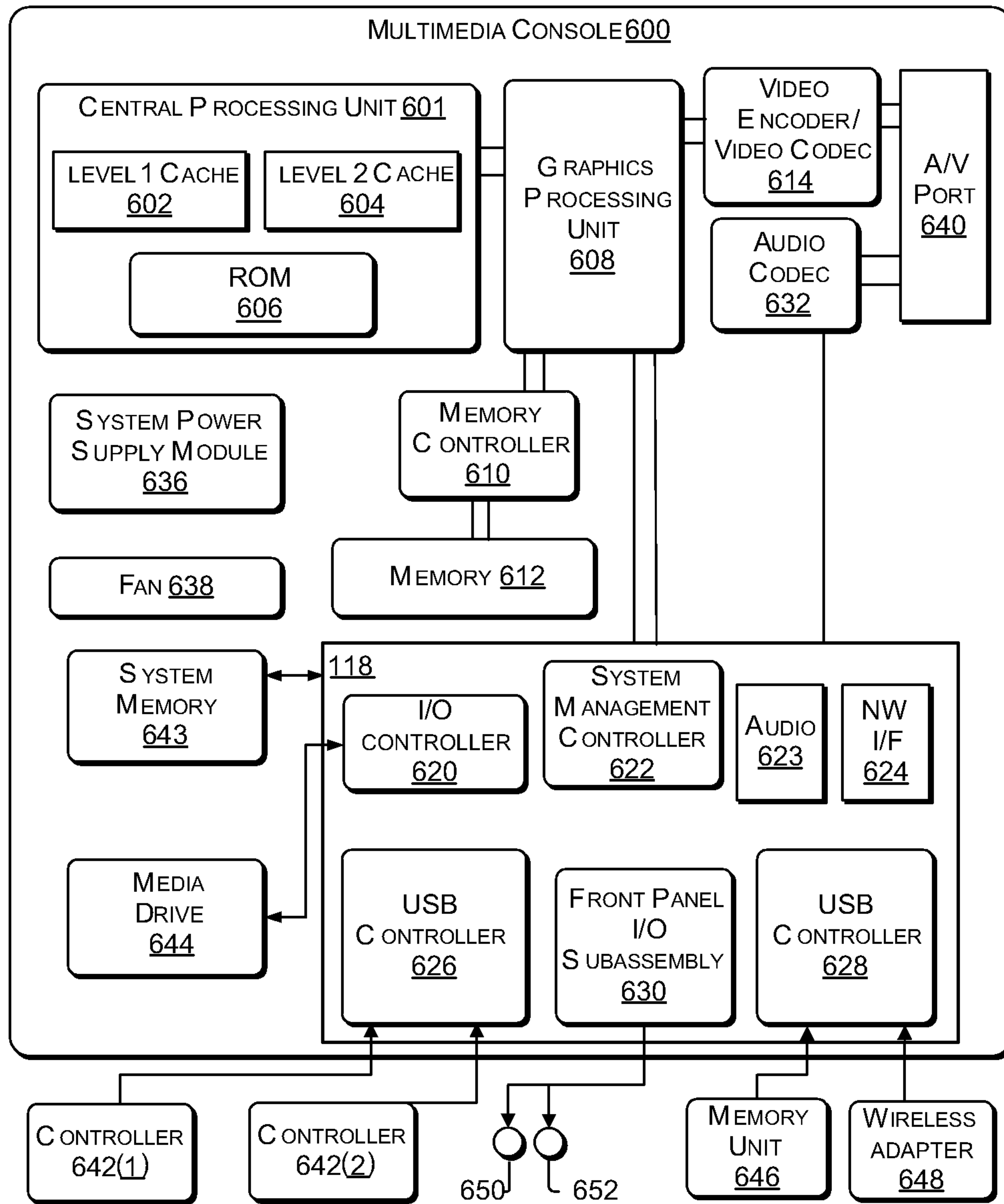


Fig. 19A

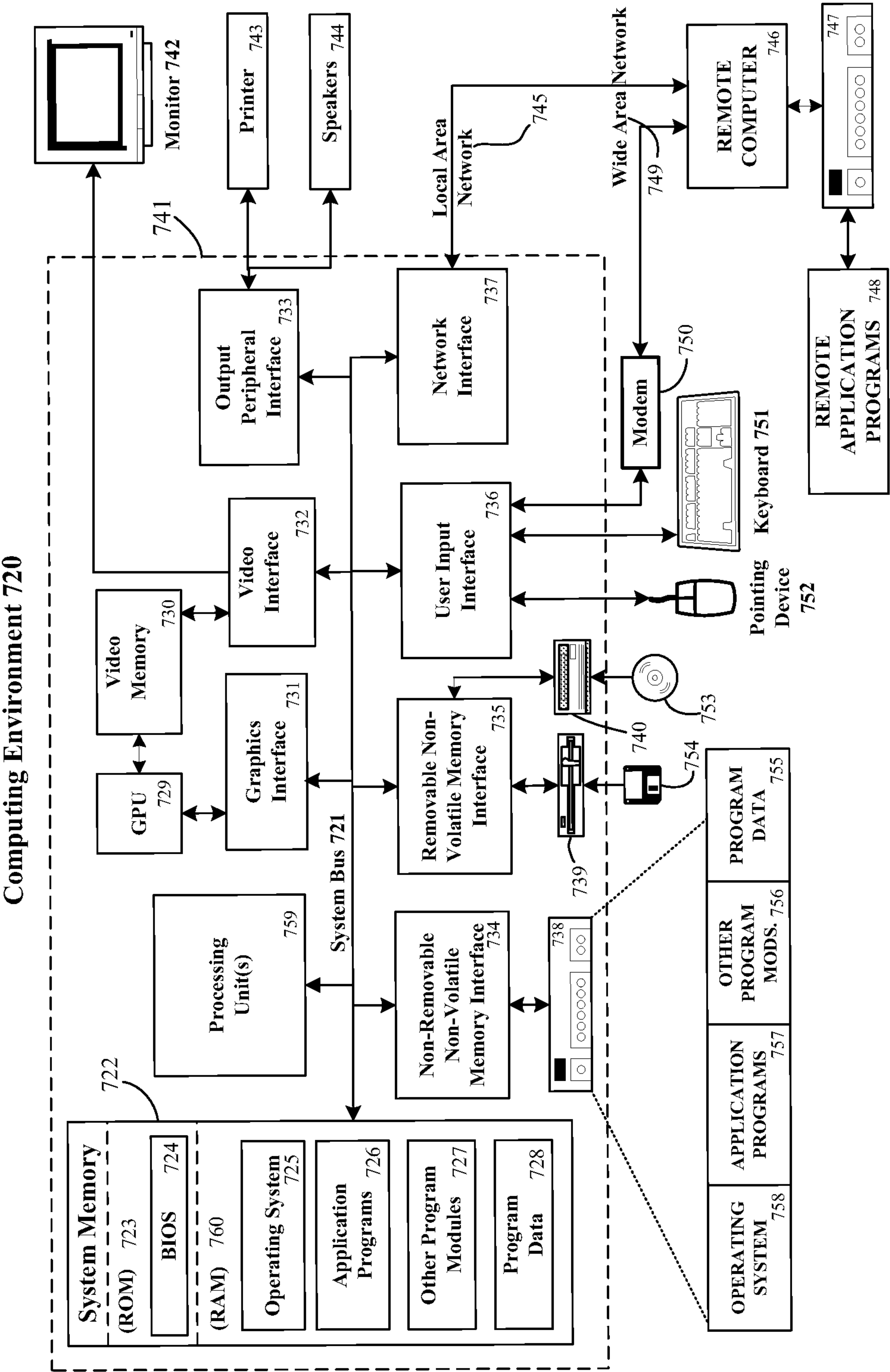


Fig. 19B

