US 20240289973A1

(19) **United States**

(12) **Patent Application Publication**

Liu et al.

(10) **Pub. No.: US 2024/0289973 A1**

(43) **Pub. Date: Aug. 29, 2024**

(54) **SYSTEMS AND METHODS FOR AN ENVIRONMENT-AWARE PREDICTIVE MODELING FRAMEWORK FOR HUMAN-ROBOT SYMBIOTIC WALKING**

(71) Applicant: **Arizona Board of Regents on Behalf of Arizona State University**, Tempe, AZ (US)

(72) Inventors: **Xiao Liu**, Tempe, AZ (US); **Geoffrey Clark**, Phoenix, AZ (US); **Heni Ben Amor**, Tempe, AZ (US)

(73) Assignee: **Arizona Board of Regents on Behalf of Arizona State University**, Temep, AZ (US)

(21) Appl. No.: **18/570,521**

(22) PCT Filed: **Jun. 14, 2022**

(86) PCT No.: **PCT/US22/33464**

§ 371 (c)(1),
(2) Date: **Dec. 14, 2023**

Related U.S. Application Data

(60) Provisional application No. 63/210,187, filed on Jun. 14, 2021.

Publication Classification

(51) **Int. Cl.**

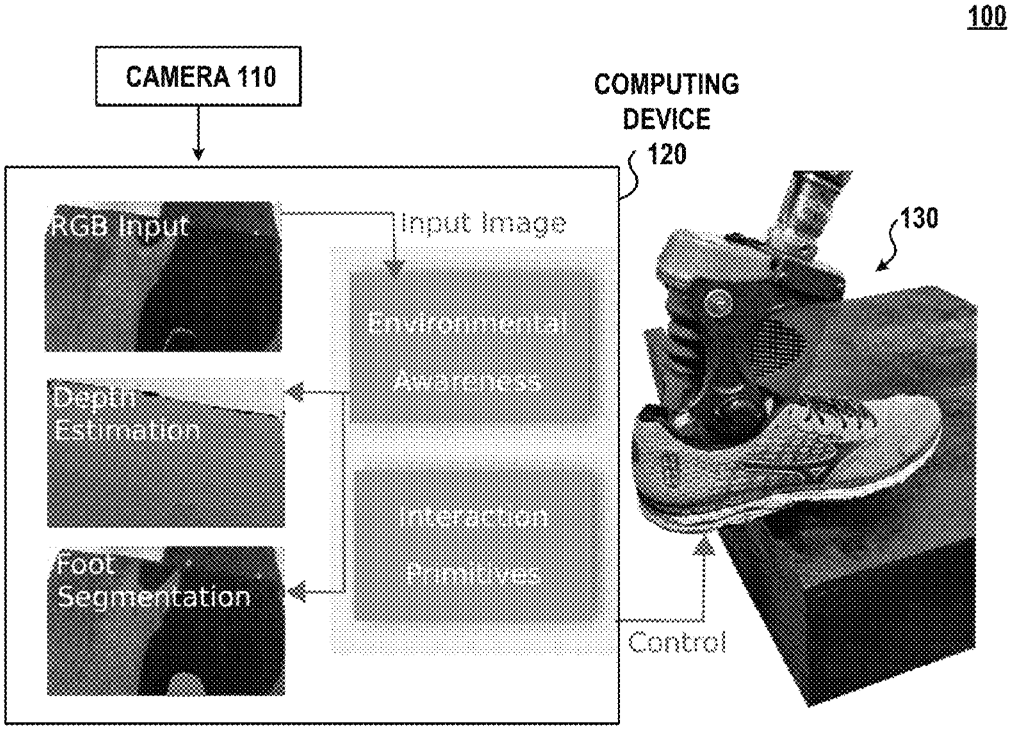
G06T 7/55	(2006.01)
A61F 2/66	(2006.01)
A61F 2/70	(2006.01)
A61H 1/02	(2006.01)
A61H 3/00	(2006.01)
G06T 7/11	(2006.01)

(52) **U.S. Cl.**

CPC ***G06T 7/55*** (2017.01); ***A61F 2/6607*** (2013.01); ***A61H 1/0266*** (2013.01); ***A61H 3/00*** (2013.01); ***G06T 7/11*** (2017.01); *A61F 2002/704* (2013.01); *G06T 2207/10024* (2013.01); *G06T 2207/20016* (2013.01); *G06T 2207/20084* (2013.01)

(57) **ABSTRACT**

An environment-aware prediction and control framework, which incorporates learned environment and terrain features into a predictive model for human-robot symbiotic walking, is disclosed herein. First, a compact deep neural network is introduced for accurate and efficient prediction of pixel-level depth maps from RGB inputs. In turn, this methodology reduces the size, weight, and cost of the necessary hardware, while adding key features such as close-range sensing, filtering, and temporal consistency. In combination with human kinematics data and demonstrated walking gaits, the extracted visual features of the environment are used to learn a probabilistic model coupling perceptions to optimal actions. The resulting data-driven controllers. Bayesian Interaction Primitives, can be used to infer in real-time optimal control actions for a lower-limb prosthesis. The inferred actions naturally take the current state of the environment and the user into account during walking.



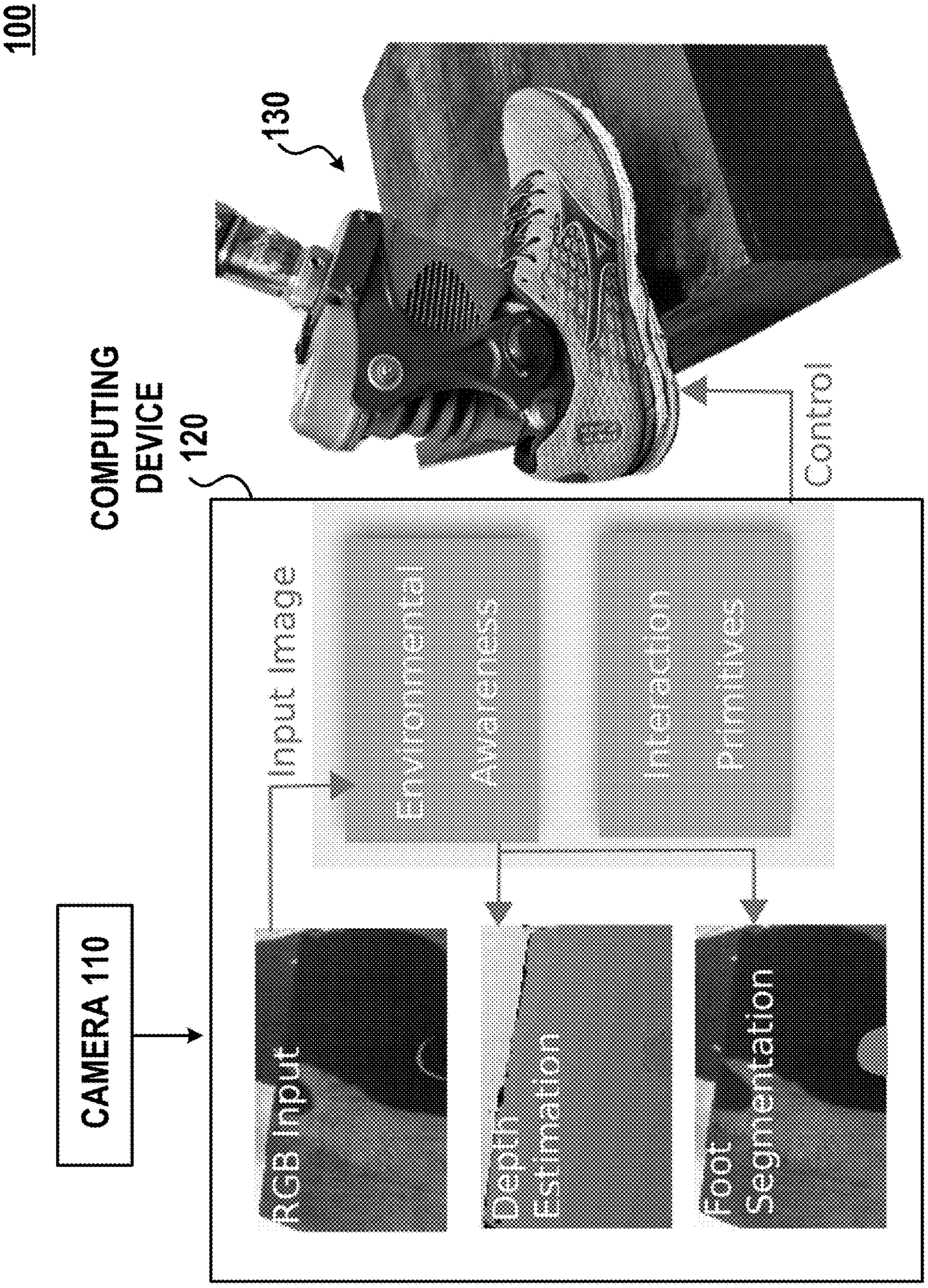

100
CAMERA 110
COMPUTING DEVICE 120
130
RGB Input
Input Image
Environmental Awareness
Interaction Primitives
Depth Estimation
Foot Segmentation
Control

FIG. 1

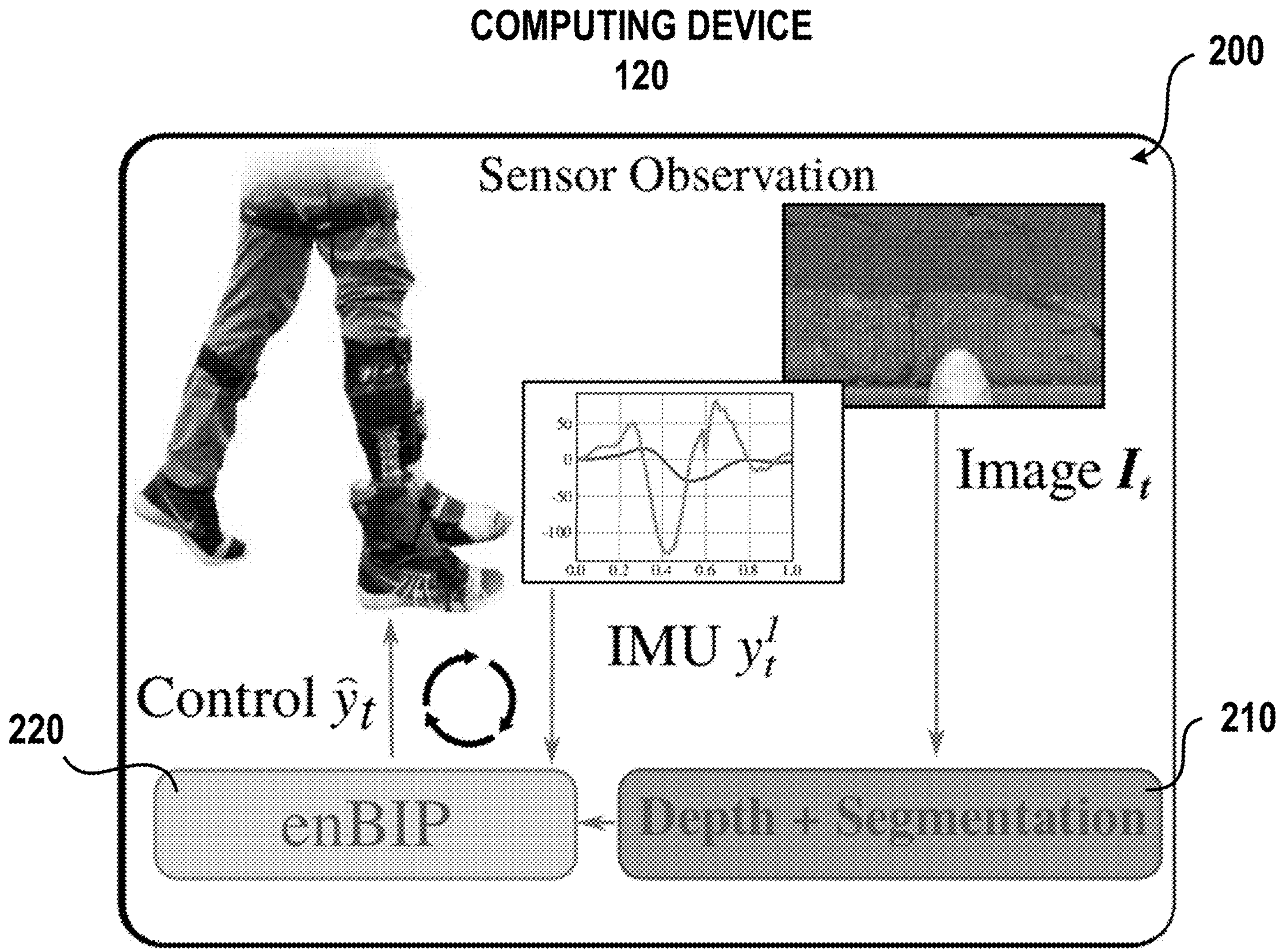

COMPUTING DEVICE
120
200
Sensor Observation
50
0
-50
-100
0.0
0.2
0.4
0.6
0.8
1.0
Image I_t
IMU y_t^l
Control $\hat{y}_t$
220
210
enBIP
Depth + Segmentation

FIG. 2A

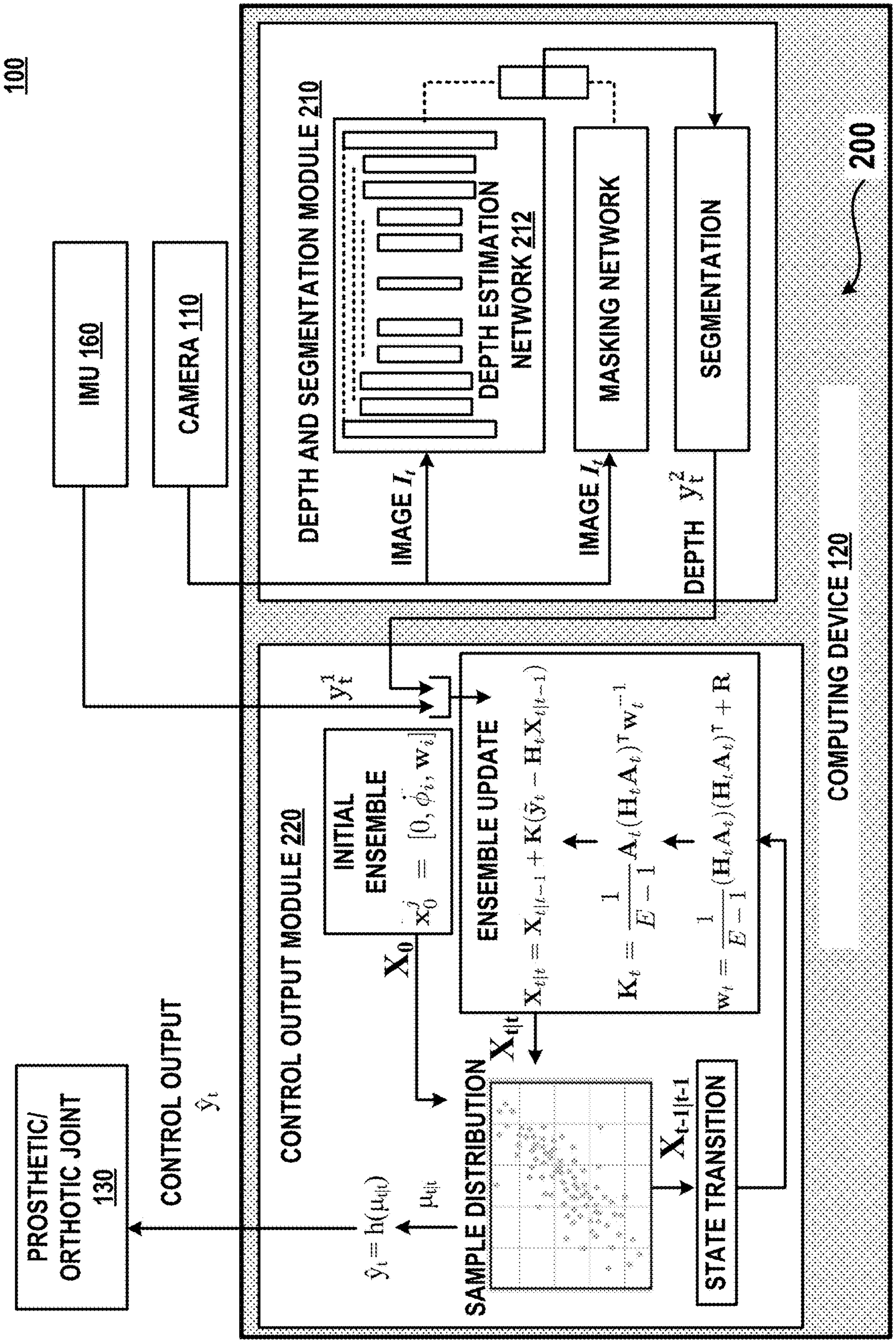

100
IMU 160
CAMERA 110
DEPTH AND SEGMENTATION MODULE 210
DEPTH ESTIMATION NETWORK 212
MASKING NETWORK
SEGMENTATION
IMAGE I_t
IMAGE I_t
DEPTH y_t^2
200
COMPUTING DEVICE 120
CONTROL OUTPUT MODULE 220
INITIAL ENSEMBLE
$x_0^j = [0, \dot{\phi}_i, w_i]$
X_0
y_t^1
ENSEMBLE UPDATE
$X_{t|t} = X_{t|t-1} + K(\tilde{y}_t - H_t X_{t|t-1})$
$K_t = \frac{1}{E-1} A_t (H_t A_t)^T w_t^{-1}$
$w_t = \frac{1}{E-1} (H_t A_t)(H_t A_t)^T + R$
$X_{t|t}$
SAMPLE DISTRIBUTION
$X_{t-1|t-1}$
STATE TRANSITION
$\mu_{t|t}$
$\hat{y}_t = h(\mu_{t|t})$
CONTROL OUTPUT $\hat{y}_t$
PROSTHETIC/ ORTHOTIC JOINT 130

FIG. 2B

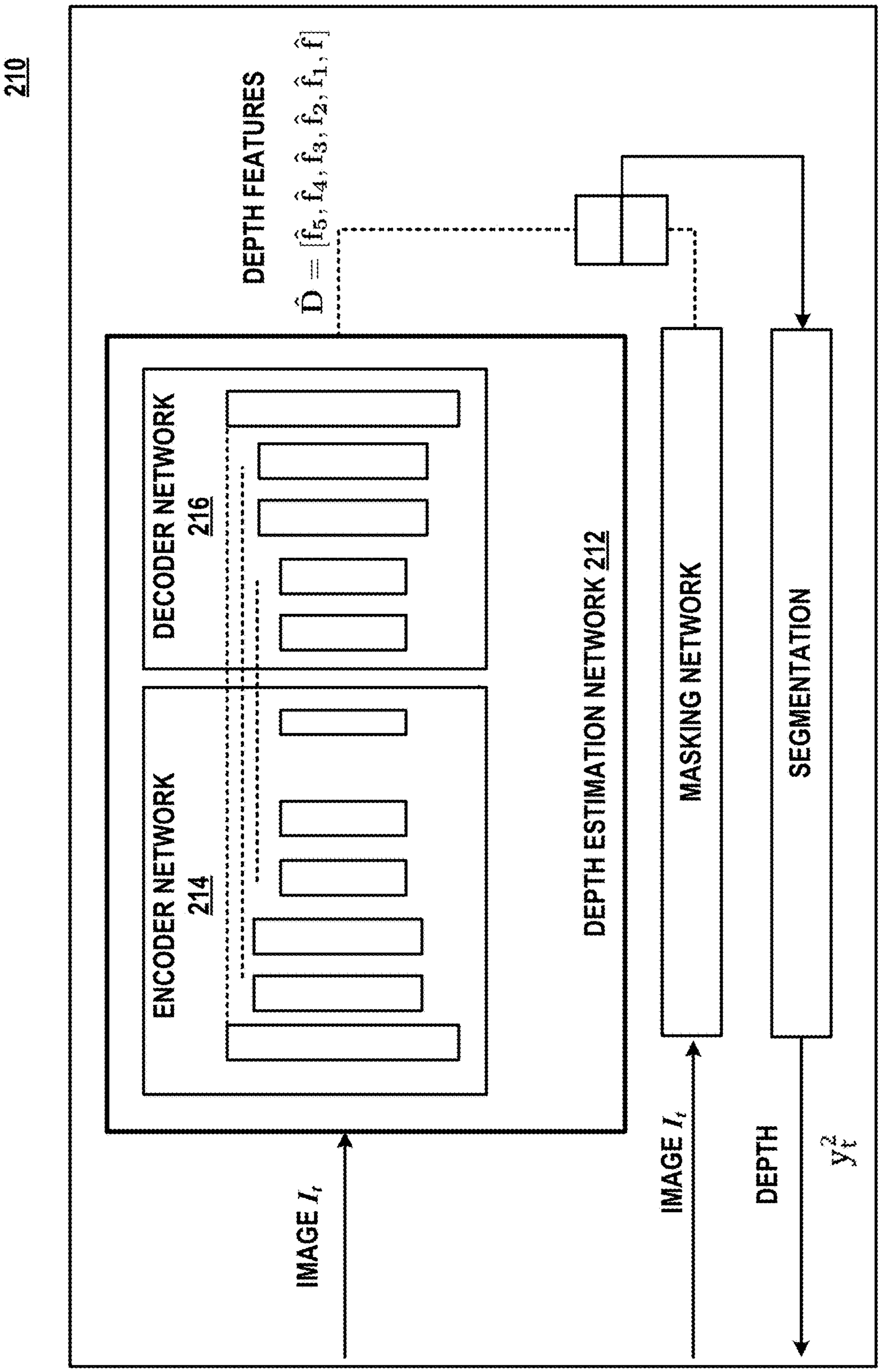

210
DEPTH FEATURES
$\hat{\mathrm{D}} = [\hat{\mathrm{f}}_5, \hat{\mathrm{f}}_4, \hat{\mathrm{f}}_3, \hat{\mathrm{f}}_2, \hat{\mathrm{f}}_1, \hat{\mathrm{f}}]$
DECODER NETWORK
216
ENCODER NETWORK
214
DEPTH ESTIMATION NETWORK 212
MASKING NETWORK
SEGMENTATION
IMAGE I_t
IMAGE I_t
DEPTH
y_t^2

FIG. 2C

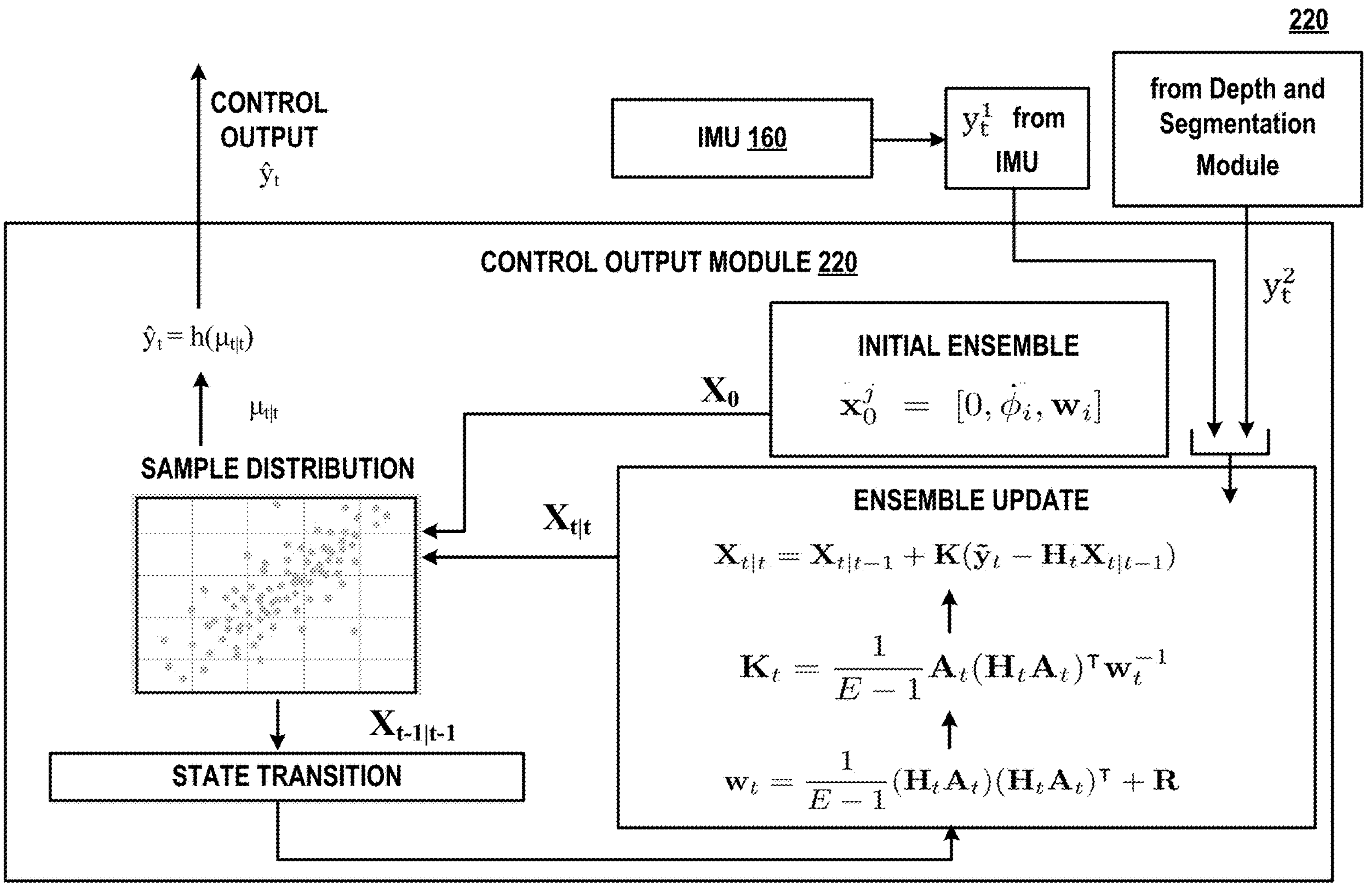

220
from Depth and Segmentation Module
IMU 160
y_t^1 from IMU
y_t^2
CONTROL OUTPUT
$\hat{y}_t$
CONTROL OUTPUT MODULE 220
$\hat{y}_t = h(\mu_{t|t})$
$\mu_{t|t}$
INITIAL ENSEMBLE
$\mathbf{x}_0^j = [0, \dot{\phi}_i, \mathbf{w}_i]$
$\mathbf{X}_0$
SAMPLE DISTRIBUTION
$\mathbf{X}_{t|t}$
ENSEMBLE UPDATE
$\mathbf{X}_{t|t} = \mathbf{X}_{t|t-1} + \mathbf{K}(\tilde{\mathbf{y}}_t - \mathbf{H}_t\mathbf{X}_{t|t-1})$
$\mathbf{K}_t = \frac{1}{E-1}\mathbf{A}_t(\mathbf{H}_t\mathbf{A}_t)^\intercal \mathbf{w}_t^{-1}$
$\mathbf{w}_t = \frac{1}{E-1}(\mathbf{H}_t\mathbf{A}_t)(\mathbf{H}_t\mathbf{A}_t)^\intercal + \mathbf{R}$
$\mathbf{X}_{t-1|t-1}$
STATE TRANSITION

FIG. 2D

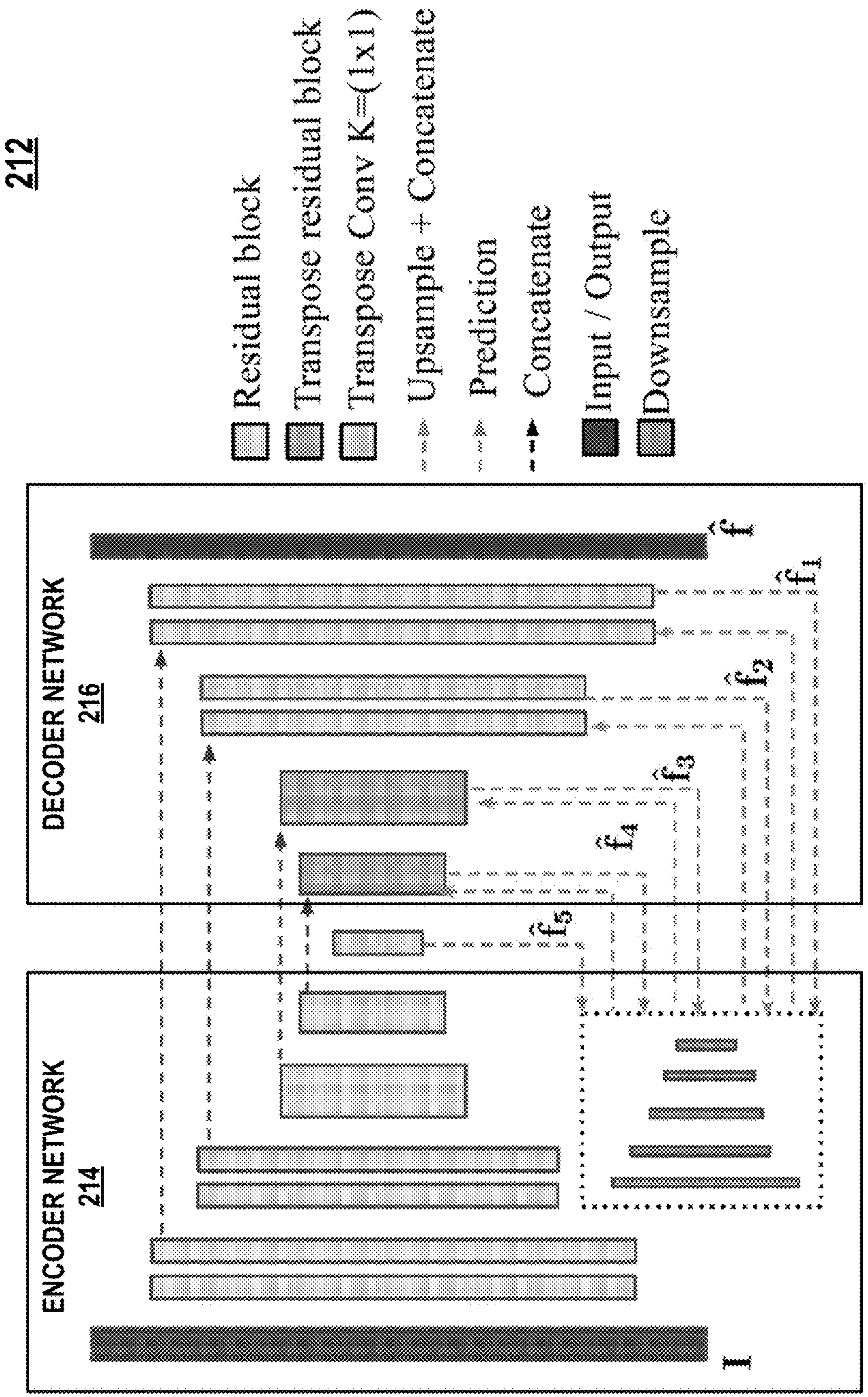

212
▨ Residual block
▨ Transpose residual block
▨ Transpose Conv K=(1x1)
Upsample + Concatenate
Prediction
Concatenate
■ Input / Output
▨ Downsample
DECODER NETWORK
216
ENCODER NETWORK
214
$\hat{f}$
$\hat{f}_1$
$\hat{f}_2$
$\hat{f}_3$
$\hat{f}_4$
$\hat{f}_5$
I

FIG. 3

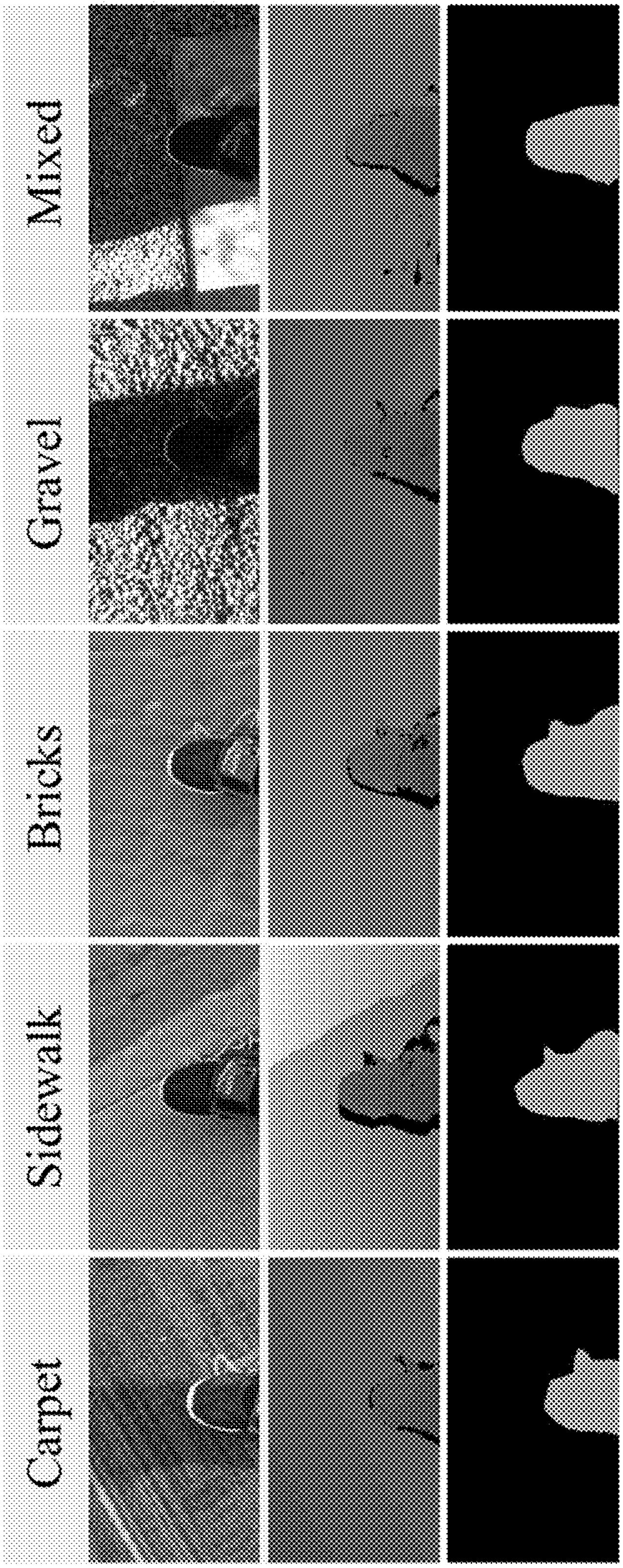


FIG. 4A

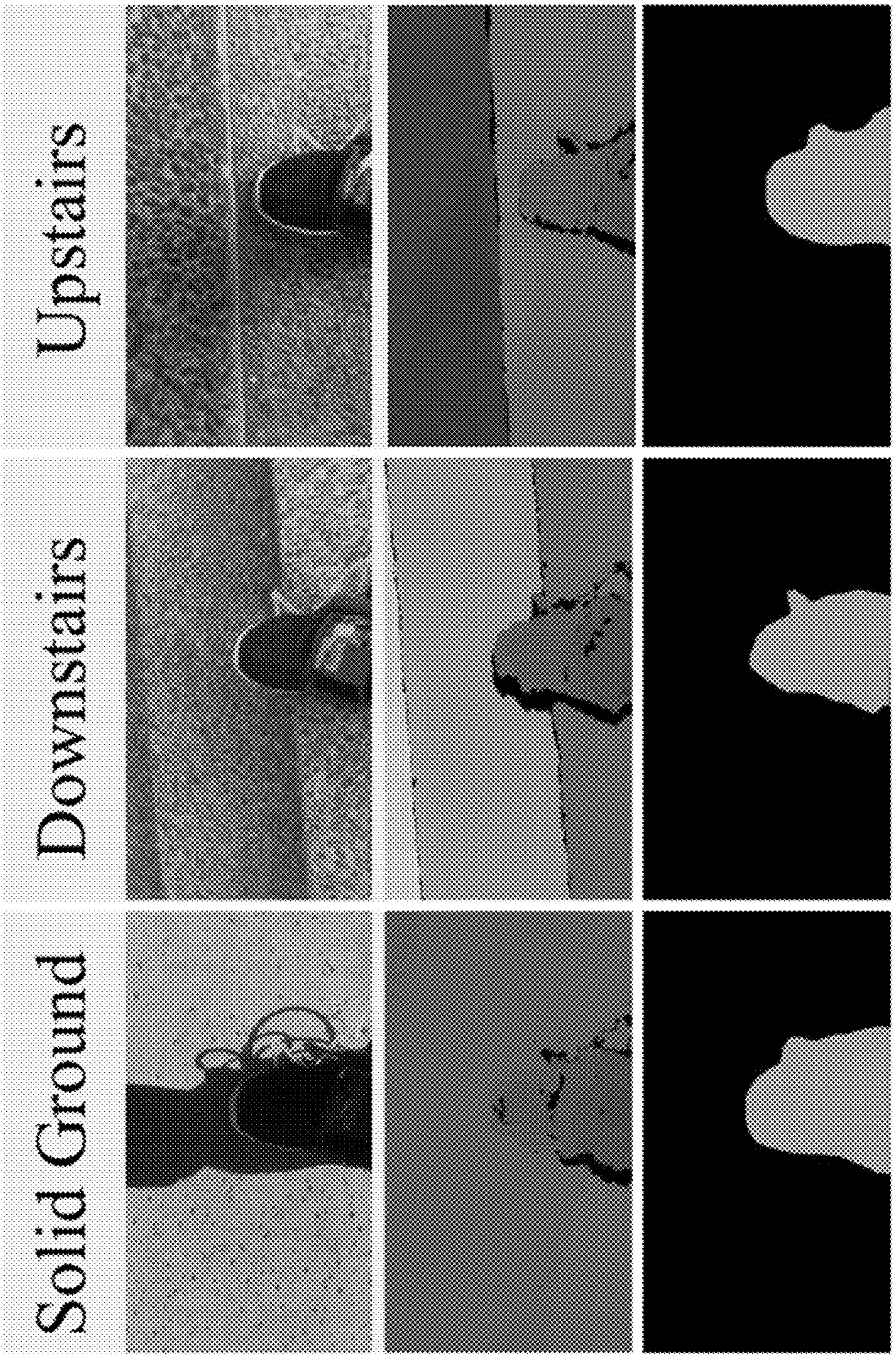


FIG. 4B

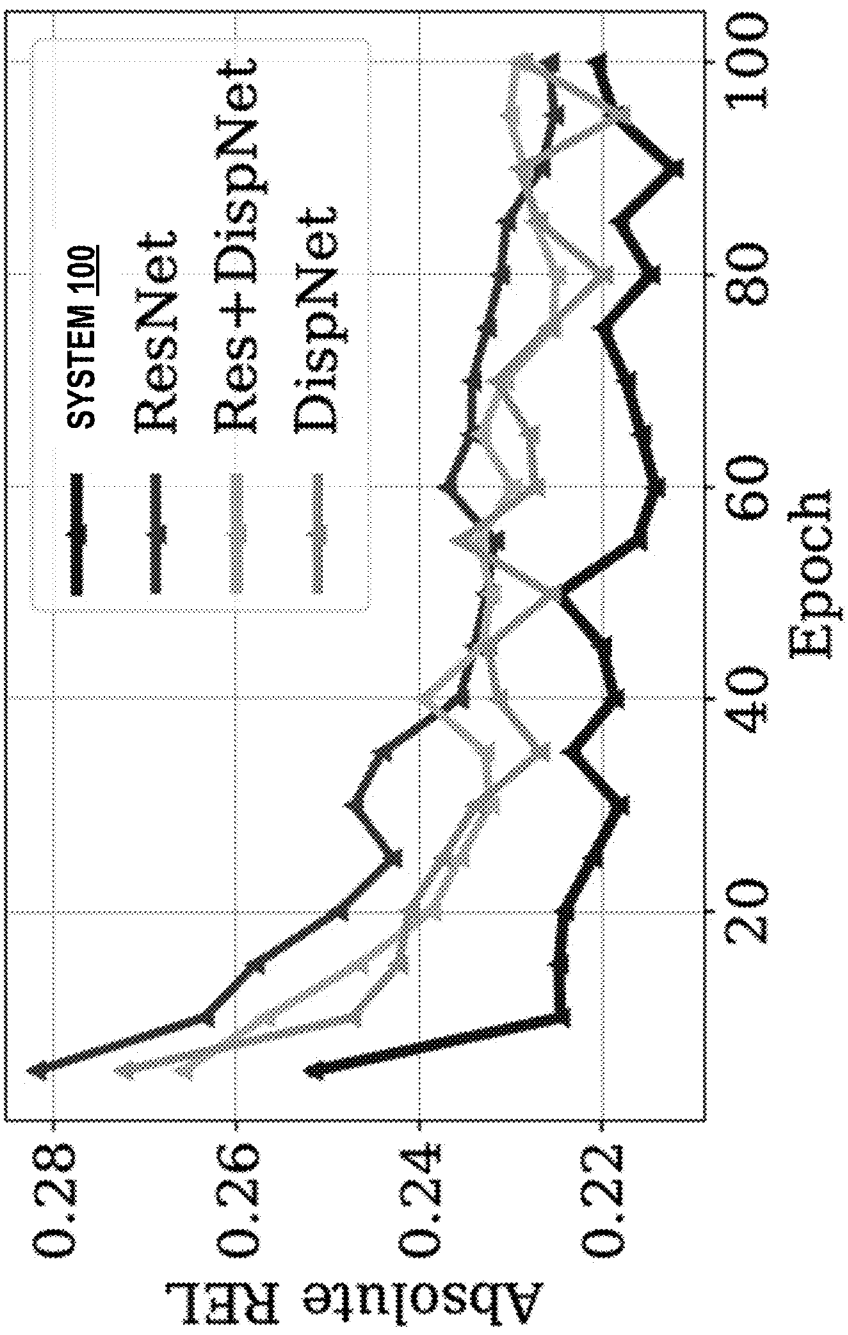

SYSTEM 100
ResNet
Res+DispNet
DispNet
Absolute REL
0.28
0.26
0.24
0.22
Epoch
20
40
60
80
100

FIG. 5A

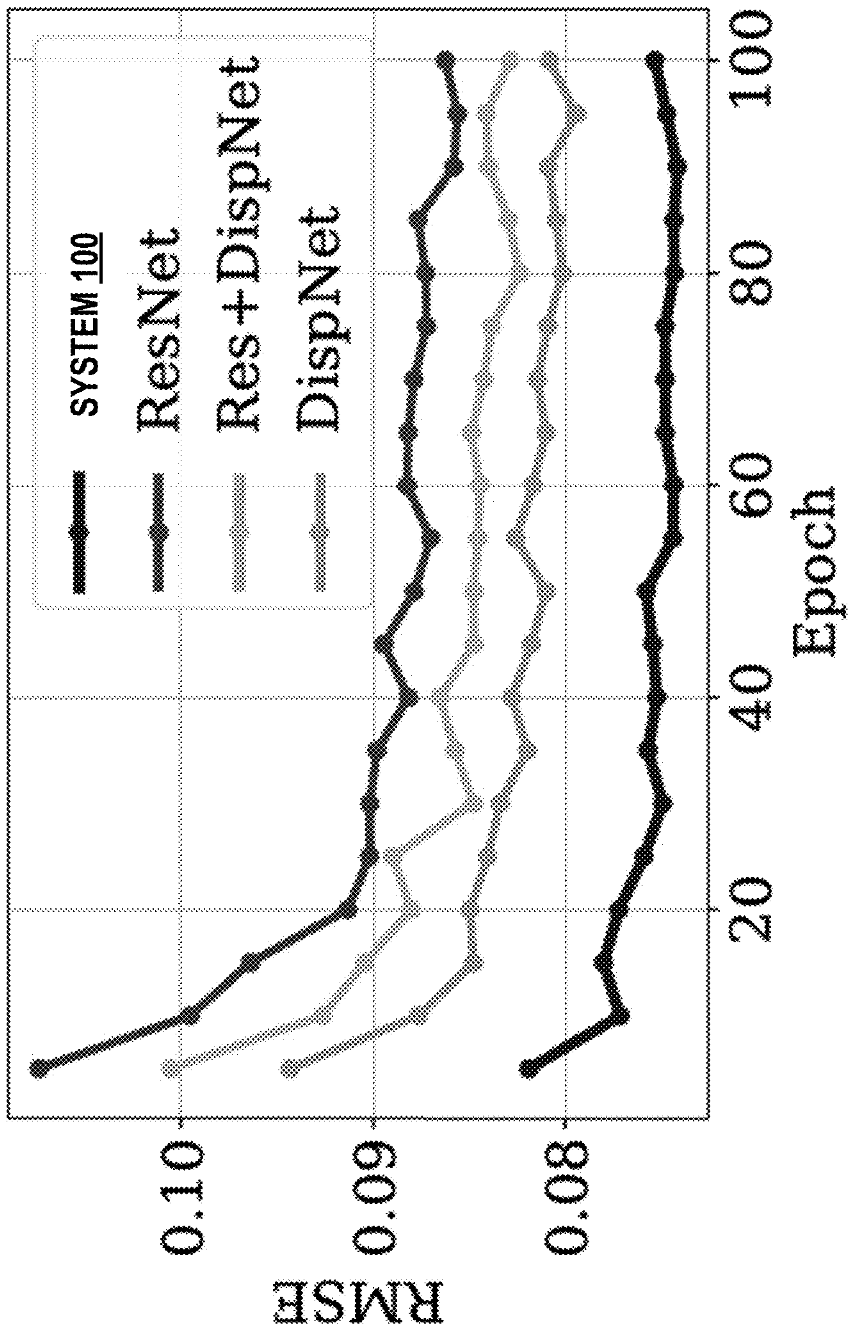

RMSE
0.10
0.09
0.08
20
40
60
80
100
Epoch
SYSTEM 100
ResNet
Res+DispNet
DispNet

FIG. 5B

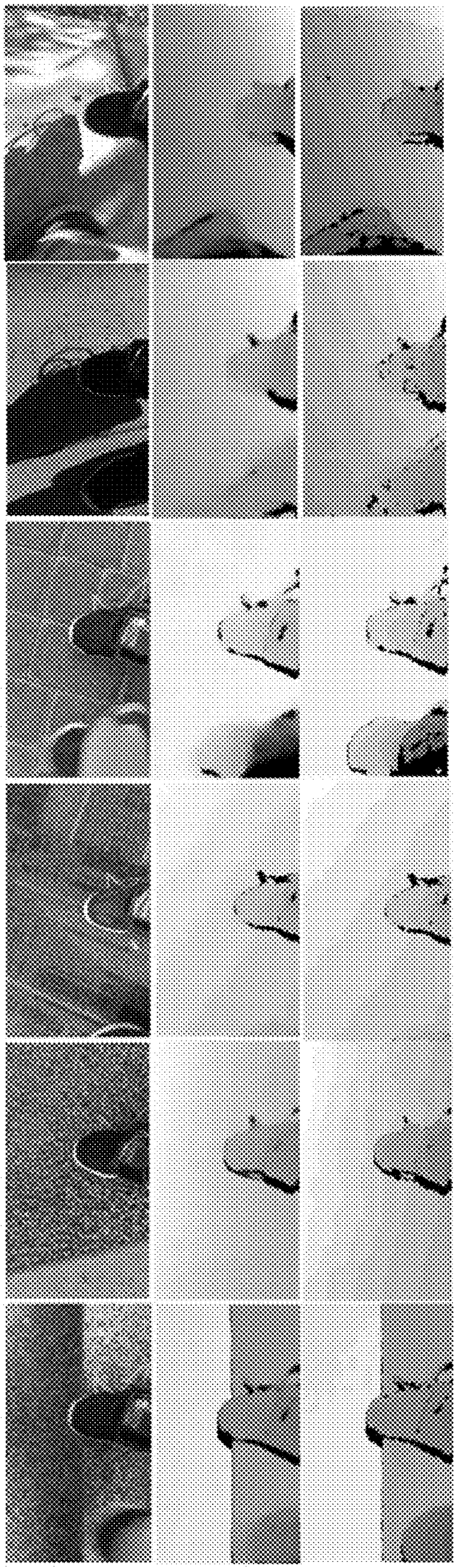


FIG. 6

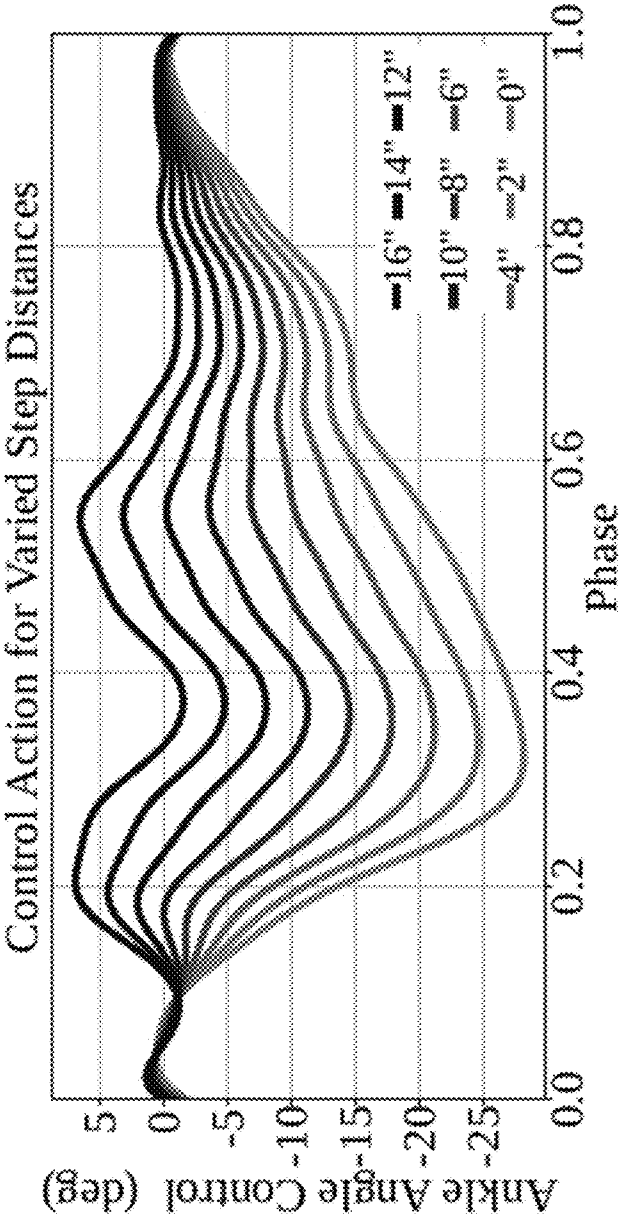

Control Action for Varied Step Distances
Ankle Angle Control (deg)
5
0
-5
-10
-15
-20
-25
0.0
0.2
0.4
0.6
0.8
1.0
Phase
16"
14"
12"
10"
8"
6"
4"
2"
0"

FIG. 7

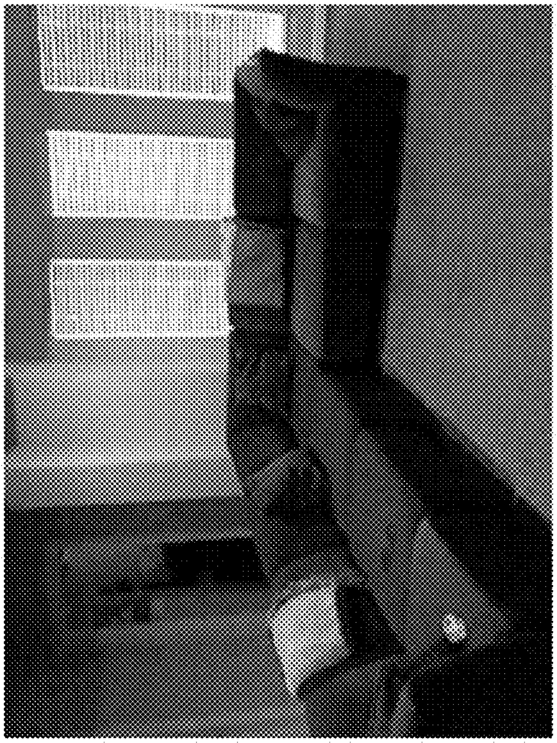


FIG. 8A

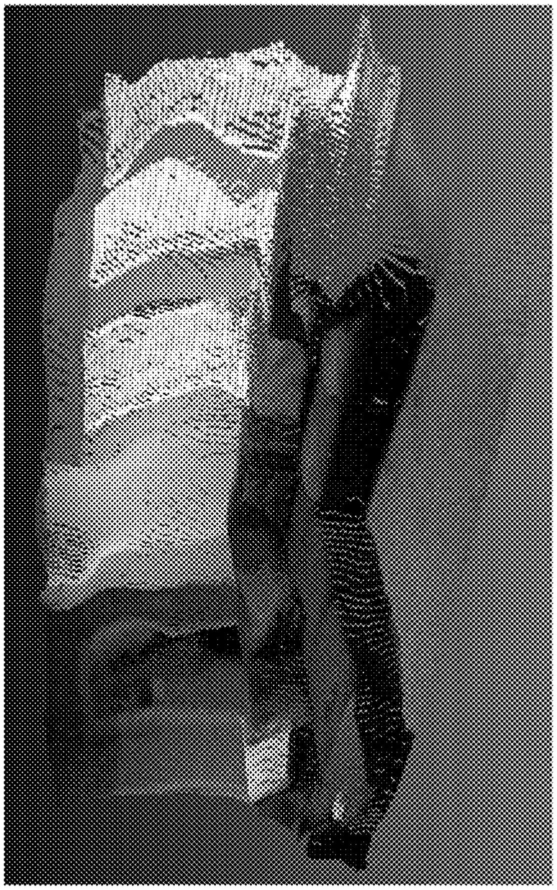


FIG. 8B

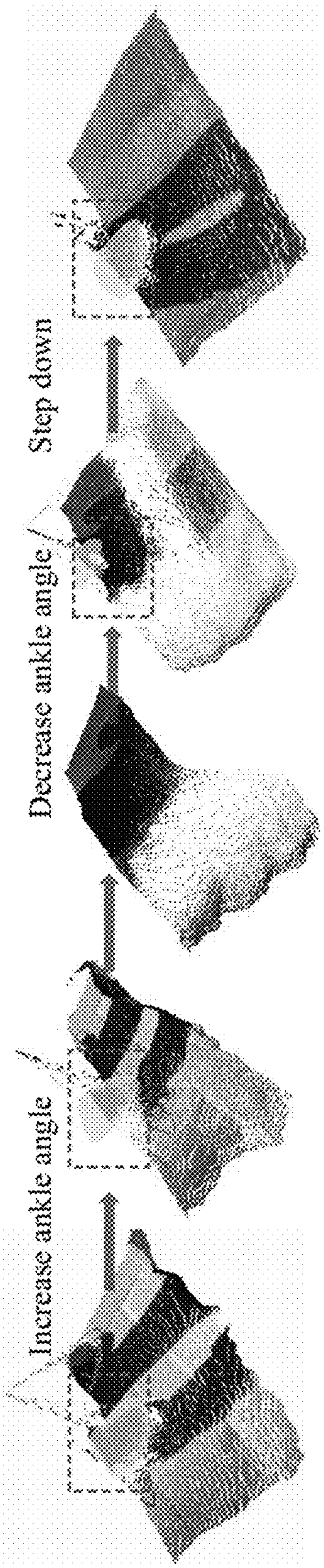


FIG. 9

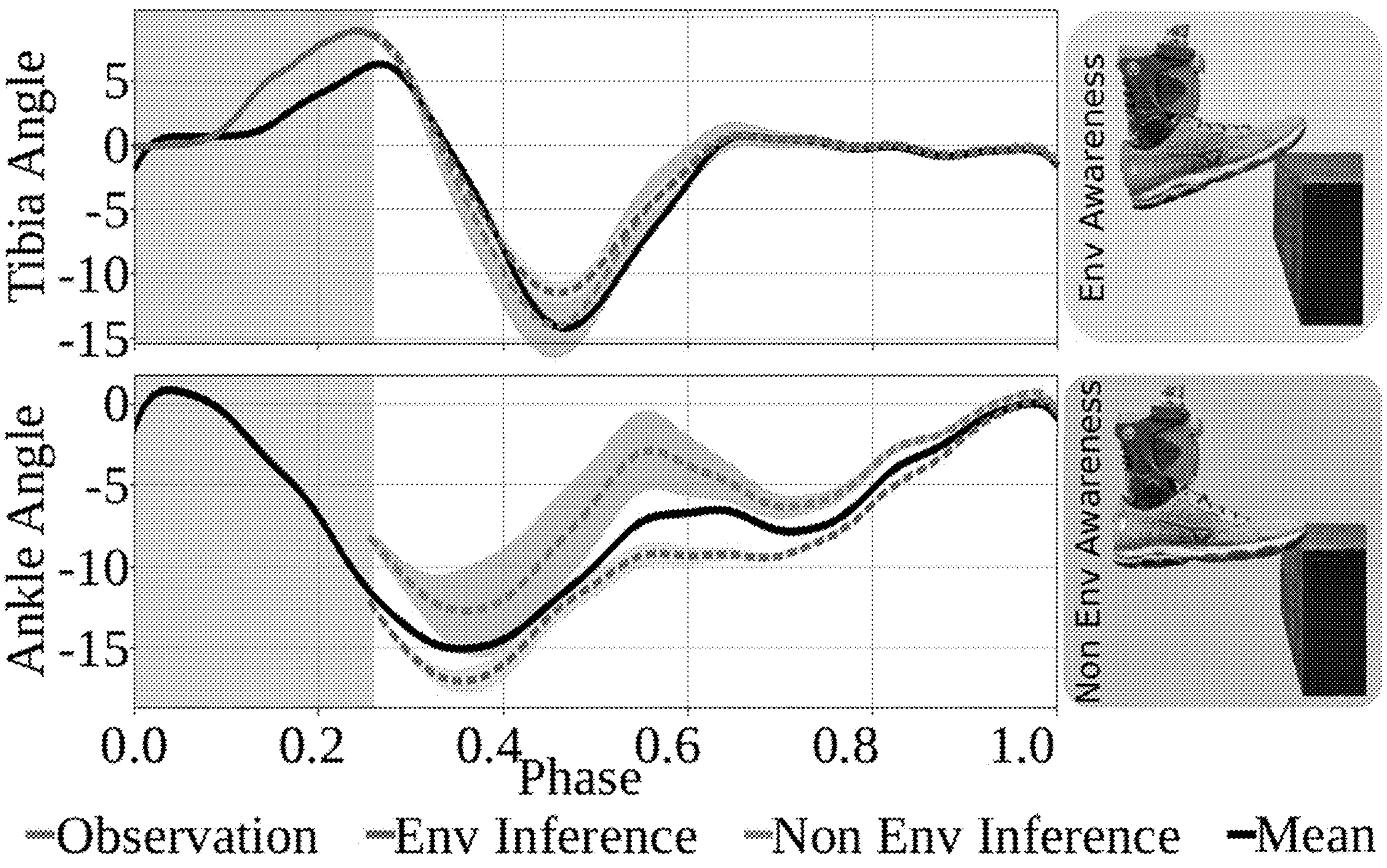

Tibia Angle
5
0
-5
-10
-15
Ankle Angle
0
-5
-10
-15
0.0
0.2
0.4
0.6
0.8
1.0
Phase
Env Awareness
Non Env Awareness
Observation
Env Inference
Non Env Inference
Mean

FIG. 10

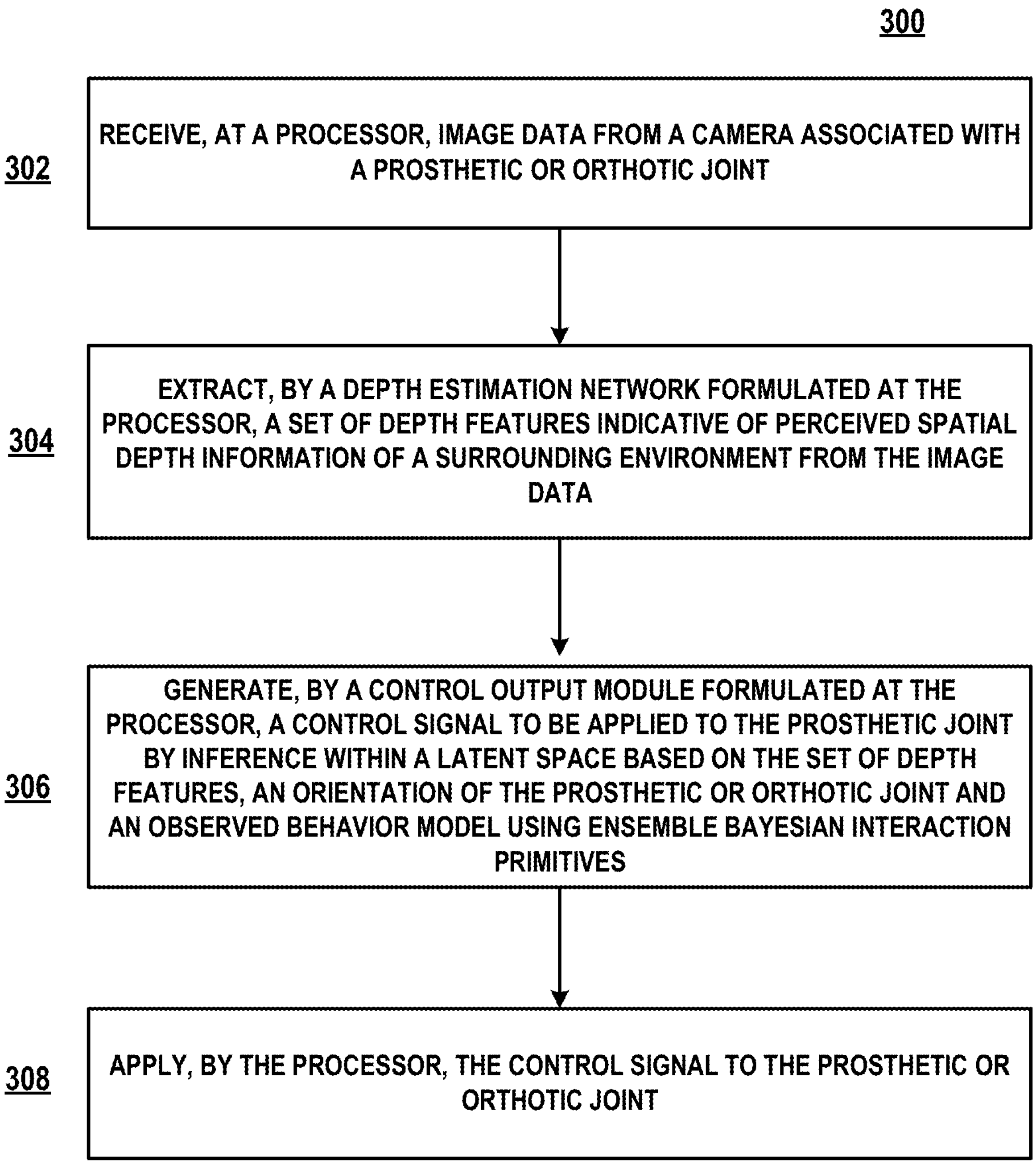

300
302
RECEIVE, AT A PROCESSOR, IMAGE DATA FROM A CAMERA ASSOCIATED WITH A PROSTHETIC OR ORTHOTIC JOINT
304
EXTRACT, BY A DEPTH ESTIMATION NETWORK FORMULATED AT THE PROCESSOR, A SET OF DEPTH FEATURES INDICATIVE OF PERCEIVED SPATIAL DEPTH INFORMATION OF A SURROUNDING ENVIRONMENT FROM THE IMAGE DATA
306
GENERATE, BY A CONTROL OUTPUT MODULE FORMULATED AT THE PROCESSOR, A CONTROL SIGNAL TO BE APPLIED TO THE PROSTHETIC JOINT BY INFERENCE WITHIN A LATENT SPACE BASED ON THE SET OF DEPTH FEATURES, AN ORIENTATION OF THE PROSTHETIC OR ORTHOTIC JOINT AND AN OBSERVED BEHAVIOR MODEL USING ENSEMBLE BAYESIAN INTERACTION PRIMITIVES
308
APPLY, BY THE PROCESSOR, THE CONTROL SIGNAL TO THE PROSTHETIC OR ORTHOTIC JOINT

FIG. 11A

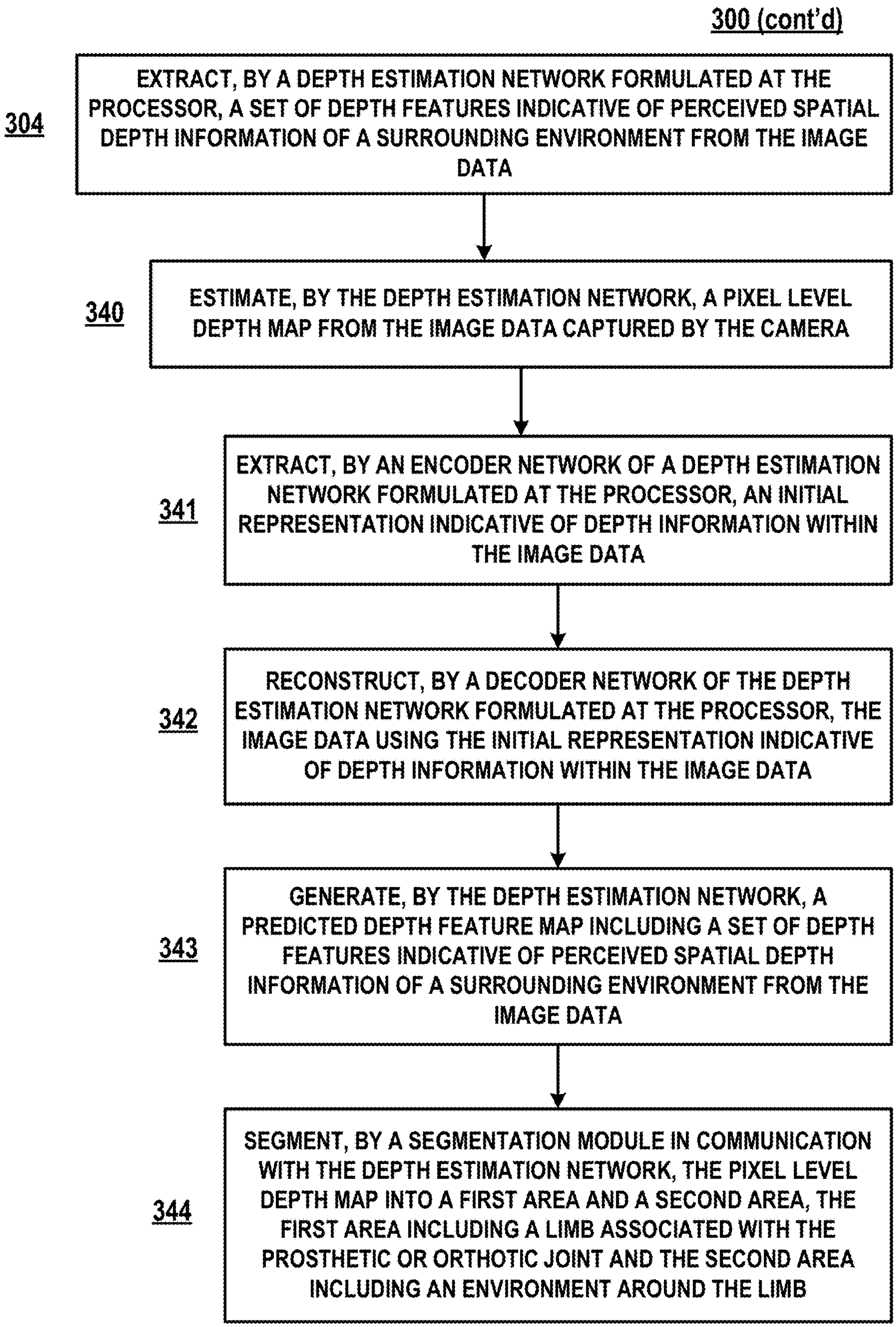

300 (cont'd)
304
EXTRACT, BY A DEPTH ESTIMATION NETWORK FORMULATED AT THE PROCESSOR, A SET OF DEPTH FEATURES INDICATIVE OF PERCEIVED SPATIAL DEPTH INFORMATION OF A SURROUNDING ENVIRONMENT FROM THE IMAGE DATA
340
ESTIMATE, BY THE DEPTH ESTIMATION NETWORK, A PIXEL LEVEL DEPTH MAP FROM THE IMAGE DATA CAPTURED BY THE CAMERA
341
EXTRACT, BY AN ENCODER NETWORK OF A DEPTH ESTIMATION NETWORK FORMULATED AT THE PROCESSOR, AN INITIAL REPRESENTATION INDICATIVE OF DEPTH INFORMATION WITHIN THE IMAGE DATA
342
RECONSTRUCT, BY A DECODER NETWORK OF THE DEPTH ESTIMATION NETWORK FORMULATED AT THE PROCESSOR, THE IMAGE DATA USING THE INITIAL REPRESENTATION INDICATIVE OF DEPTH INFORMATION WITHIN THE IMAGE DATA
343
GENERATE, BY THE DEPTH ESTIMATION NETWORK, A PREDICTED DEPTH FEATURE MAP INCLUDING A SET OF DEPTH FEATURES INDICATIVE OF PERCEIVED SPATIAL DEPTH INFORMATION OF A SURROUNDING ENVIRONMENT FROM THE IMAGE DATA
344
SEGMENT, BY A SEGMENTATION MODULE IN COMMUNICATION WITH THE DEPTH ESTIMATION NETWORK, THE PIXEL LEVEL DEPTH MAP INTO A FIRST AREA AND A SECOND AREA, THE FIRST AREA INCLUDING A LIMB ASSOCIATED WITH THE PROSTHETIC OR ORTHOTIC JOINT AND THE SECOND AREA INCLUDING AN ENVIRONMENT AROUND THE LIMB

FIG. 11B

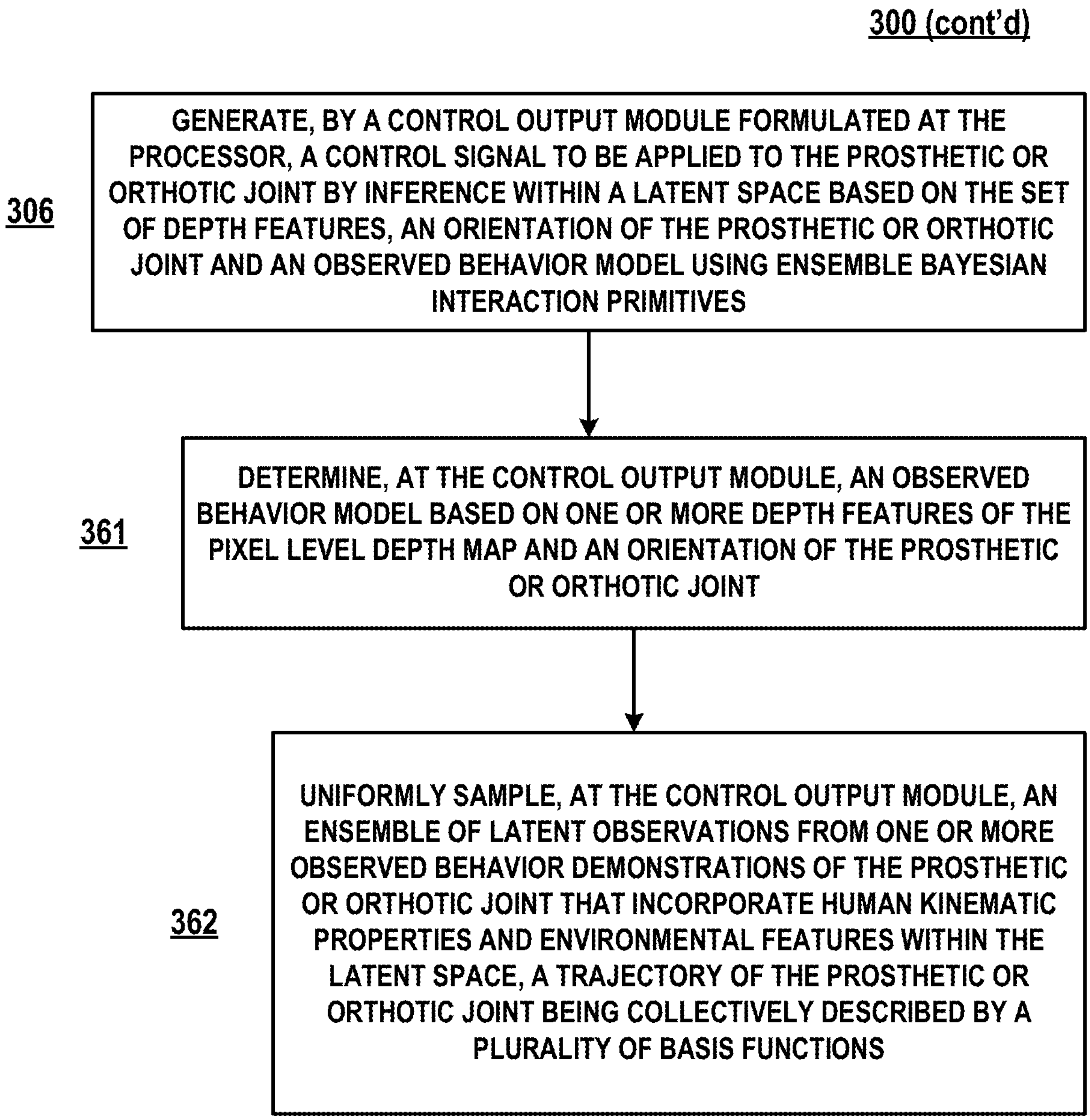

300 (cont'd)
306
GENERATE, BY A CONTROL OUTPUT MODULE FORMULATED AT THE PROCESSOR, A CONTROL SIGNAL TO BE APPLIED TO THE PROSTHETIC OR ORTHOTIC JOINT BY INFERENCE WITHIN A LATENT SPACE BASED ON THE SET OF DEPTH FEATURES, AN ORIENTATION OF THE PROSTHETIC OR ORTHOTIC JOINT AND AN OBSERVED BEHAVIOR MODEL USING ENSEMBLE BAYESIAN INTERACTION PRIMITIVES
361
DETERMINE, AT THE CONTROL OUTPUT MODULE, AN OBSERVED BEHAVIOR MODEL BASED ON ONE OR MORE DEPTH FEATURES OF THE PIXEL LEVEL DEPTH MAP AND AN ORIENTATION OF THE PROSTHETIC OR ORTHOTIC JOINT
362
UNIFORMLY SAMPLE, AT THE CONTROL OUTPUT MODULE, AN ENSEMBLE OF LATENT OBSERVATIONS FROM ONE OR MORE OBSERVED BEHAVIOR DEMONSTRATIONS OF THE PROSTHETIC OR ORTHOTIC JOINT THAT INCORPORATE HUMAN KINEMATIC PROPERTIES AND ENVIRONMENTAL FEATURES WITHIN THE LATENT SPACE, A TRAJECTORY OF THE PROSTHETIC OR ORTHOTIC JOINT BEING COLLECTIVELY DESCRIBED BY A PLURALITY OF BASIS FUNCTIONS

FIG. 11C

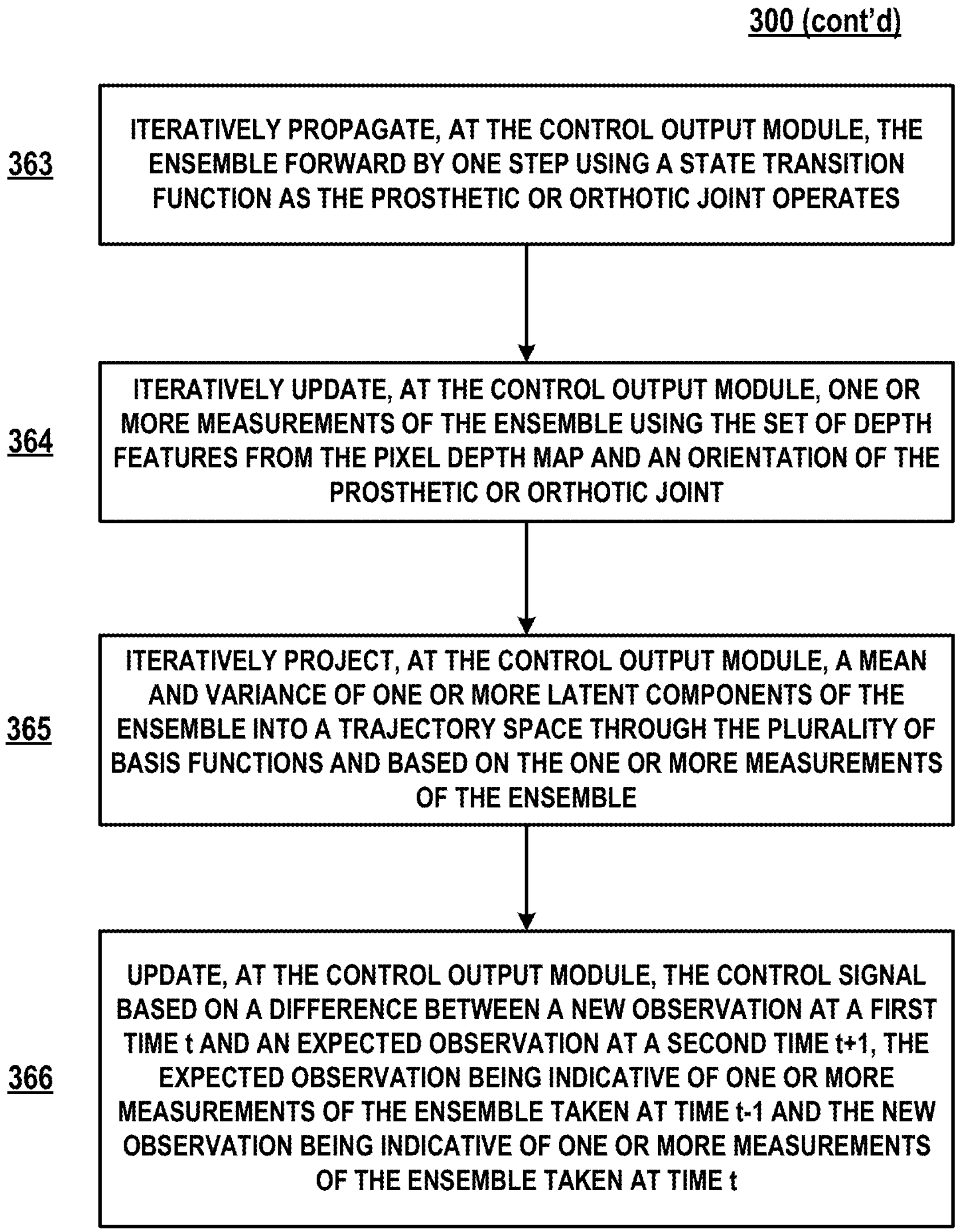

300 (cont'd)
363
ITERATIVELY PROPAGATE, AT THE CONTROL OUTPUT MODULE, THE ENSEMBLE FORWARD BY ONE STEP USING A STATE TRANSITION FUNCTION AS THE PROSTHETIC OR ORTHOTIC JOINT OPERATES
364
ITERATIVELY UPDATE, AT THE CONTROL OUTPUT MODULE, ONE OR MORE MEASUREMENTS OF THE ENSEMBLE USING THE SET OF DEPTH FEATURES FROM THE PIXEL DEPTH MAP AND AN ORIENTATION OF THE PROSTHETIC OR ORTHOTIC JOINT
365
ITERATIVELY PROJECT, AT THE CONTROL OUTPUT MODULE, A MEAN AND VARIANCE OF ONE OR MORE LATENT COMPONENTS OF THE ENSEMBLE INTO A TRAJECTORY SPACE THROUGH THE PLURALITY OF BASIS FUNCTIONS AND BASED ON THE ONE OR MORE MEASUREMENTS OF THE ENSEMBLE
366
UPDATE, AT THE CONTROL OUTPUT MODULE, THE CONTROL SIGNAL BASED ON A DIFFERENCE BETWEEN A NEW OBSERVATION AT A FIRST TIME t AND AN EXPECTED OBSERVATION AT A SECOND TIME t+1, THE EXPECTED OBSERVATION BEING INDICATIVE OF ONE OR MORE MEASUREMENTS OF THE ENSEMBLE TAKEN AT TIME t-1 AND THE NEW OBSERVATION BEING INDICATIVE OF ONE OR MORE MEASUREMENTS OF THE ENSEMBLE TAKEN AT TIME t

FIG. 11D

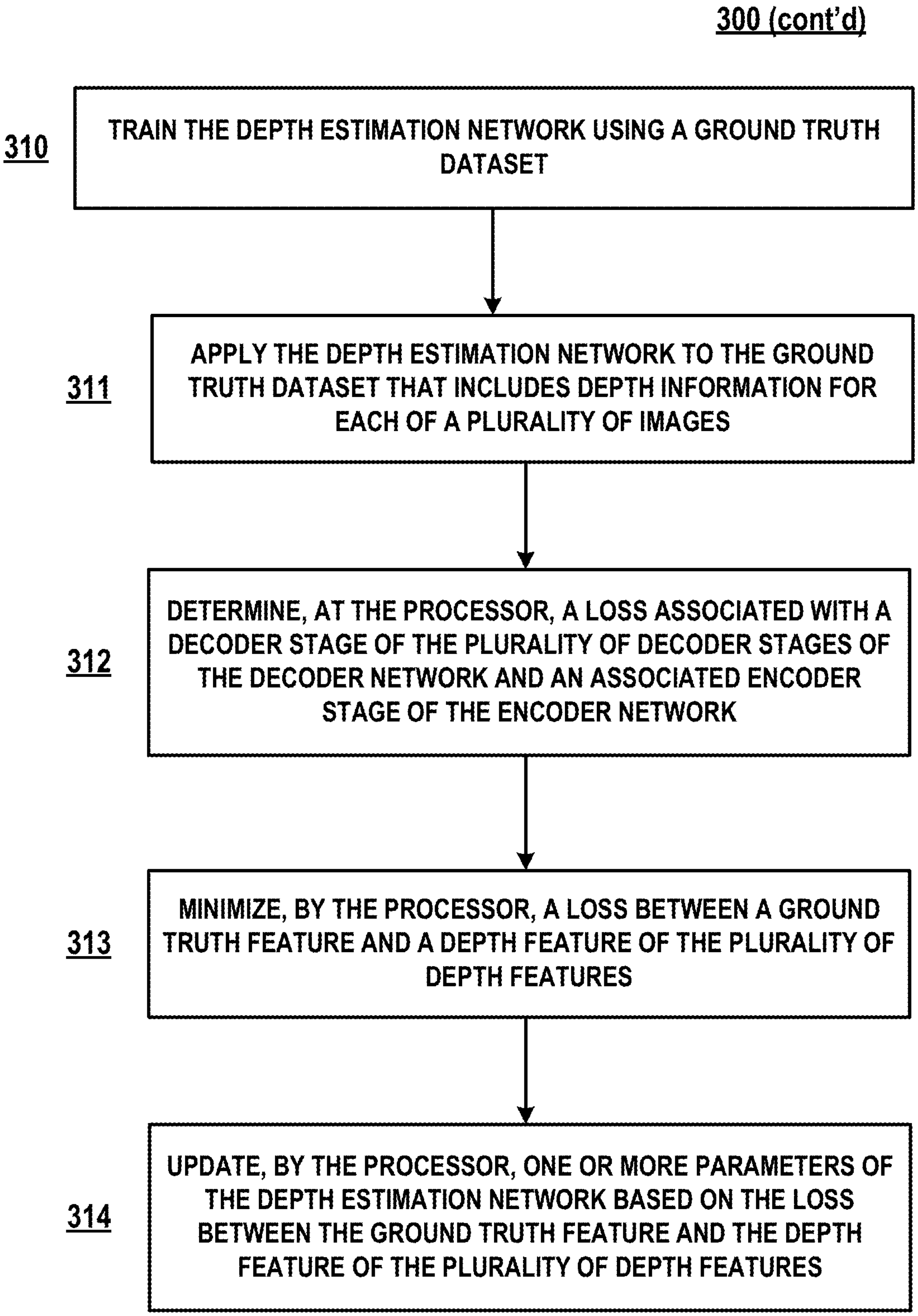

300 (cont'd)
310
TRAIN THE DEPTH ESTIMATION NETWORK USING A GROUND TRUTH DATASET
311
APPLY THE DEPTH ESTIMATION NETWORK TO THE GROUND TRUTH DATASET THAT INCLUDES DEPTH INFORMATION FOR EACH OF A PLURALITY OF IMAGES
312
DETERMINE, AT THE PROCESSOR, A LOSS ASSOCIATED WITH A DECODER STAGE OF THE PLURALITY OF DECODER STAGES OF THE DECODER NETWORK AND AN ASSOCIATED ENCODER STAGE OF THE ENCODER NETWORK
313
MINIMIZE, BY THE PROCESSOR, A LOSS BETWEEN A GROUND TRUTH FEATURE AND A DEPTH FEATURE OF THE PLURALITY OF DEPTH FEATURES
314
UPDATE, BY THE PROCESSOR, ONE OR MORE PARAMETERS OF THE DEPTH ESTIMATION NETWORK BASED ON THE LOSS BETWEEN THE GROUND TRUTH FEATURE AND THE DEPTH FEATURE OF THE PLURALITY OF DEPTH FEATURES

FIG. 11E

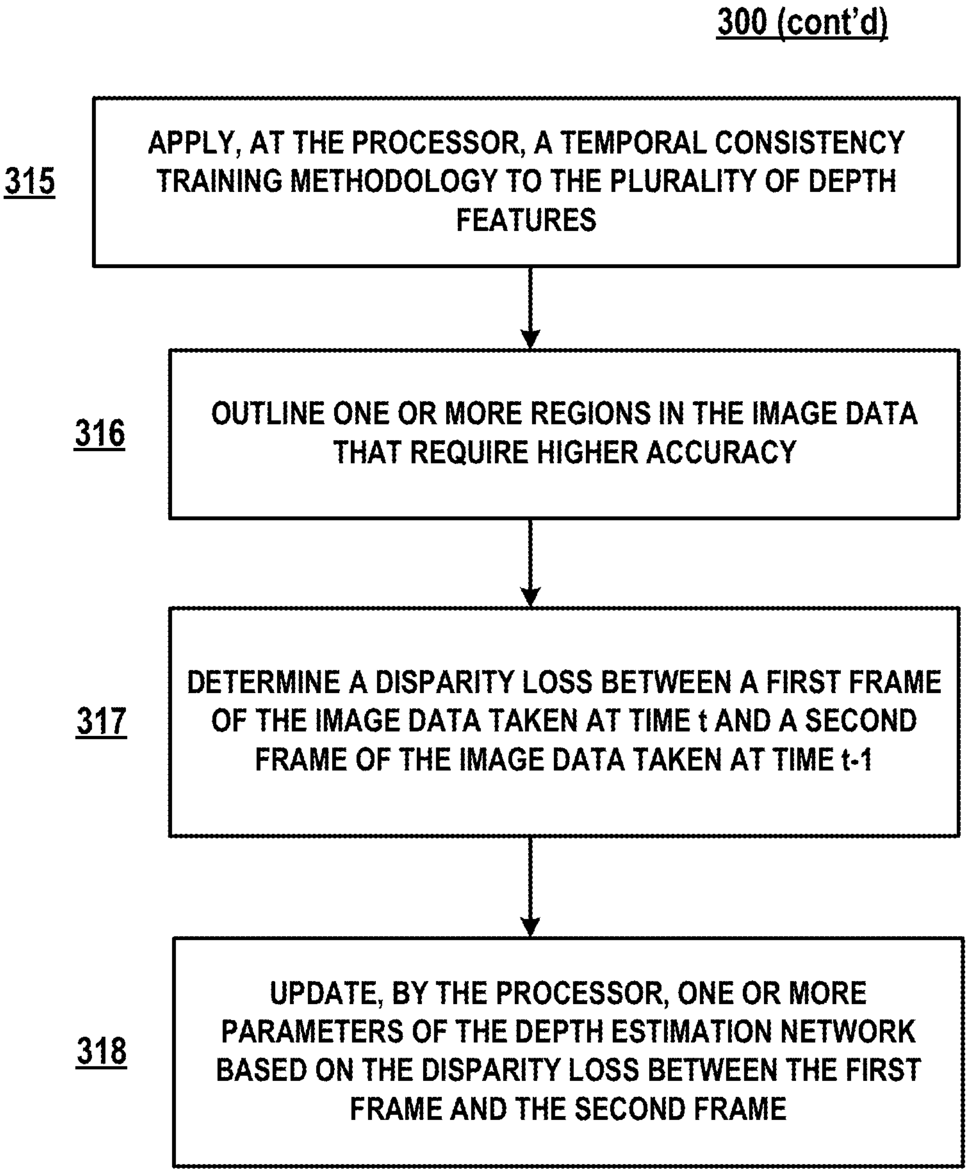

300 (cont'd)
315
APPLY, AT THE PROCESSOR, A TEMPORAL CONSISTENCY TRAINING METHODOLOGY TO THE PLURALITY OF DEPTH FEATURES
316
OUTLINE ONE OR MORE REGIONS IN THE IMAGE DATA THAT REQUIRE HIGHER ACCURACY
317
DETERMINE A DISPARITY LOSS BETWEEN A FIRST FRAME OF THE IMAGE DATA TAKEN AT TIME t AND A SECOND FRAME OF THE IMAGE DATA TAKEN AT TIME t-1
318
UPDATE, BY THE PROCESSOR, ONE OR MORE PARAMETERS OF THE DEPTH ESTIMATION NETWORK BASED ON THE DISPARITY LOSS BETWEEN THE FIRST FRAME AND THE SECOND FRAME

FIG. 11F

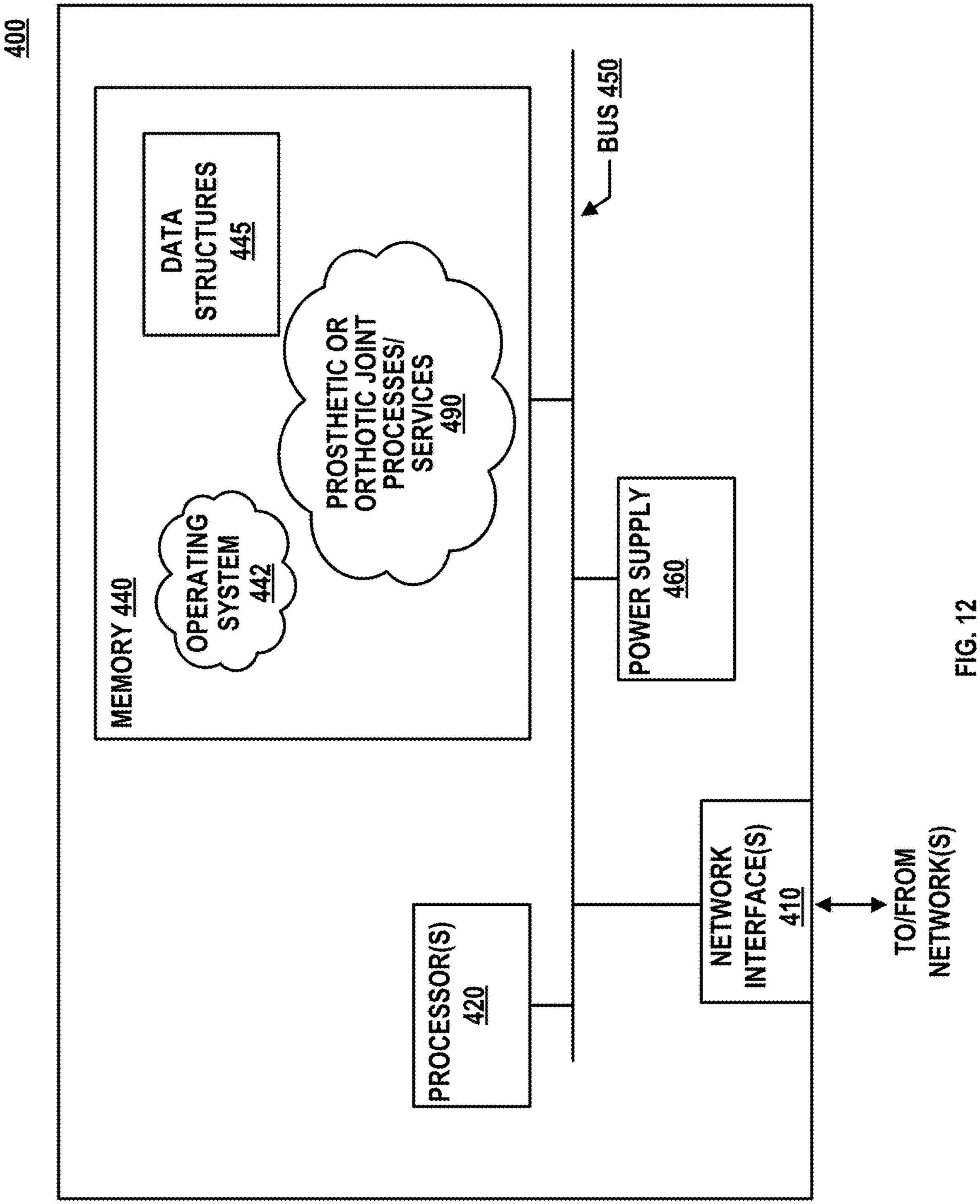

400
MEMORY 440
DATA STRUCTURES 445
OPERATING SYSTEM 442
PROSTHETIC OR ORTHOTIC JOINT PROCESSES/ SERVICES 490
BUS 450
POWER SUPPLY 460
PROCESSOR(S) 420
NETWORK INTERFACE(S) 410
TO/FROM NETWORK(S)

FIG. 12

SYSTEMS AND METHODS FOR AN ENVIRONMENT-AWARE PREDICTIVE MODELING FRAMEWORK FOR HUMAN-ROBOT SYMBIOTIC WALKING

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This is a PCT application that claims benefit to U.S. Provisional Patent Application Ser. No. 63/210,187 filed 14 Jun. 2021, which is herein incorporated by reference in its entirety.

GOVERNMENT SUPPORT

[0002] This invention was made with government support under 1749783 awarded by the National Science Foundation. The government has certain rights in the invention.

FIELD

[0003] The present disclosure generally relates to human-robot interactive systems, and in particular to a system and associated method for an environment-aware predictive modeling framework for a prosthetic or orthotic joint.

BACKGROUND

[0004] Robotic prostheses and orthotics have the potential to change the lives of millions of lower-limb amputees or non-amputees with mobility-related problems for the better by providing critical support during legged locomotion. Powered prostheses and orthotics enable complex capabilities such as level-ground walking and running or stair climbing, while also enabling reductions in metabolic cost and improvements in ergonomic comfort. However, most existing devices are tuned toward and heavily focus on unobstructed level-ground walking, to the detriment of other gait modes-especially those required in dynamic environments. Limitations to the range and adaptivity of gaits has negatively impacted the ability of amputees to navigate dynamic landscapes. Yet, the primary cause of falls is inadequate foot clearance during obstacle traversal during obstacle traversal. In many cases only millimeters decide whether a gait will be safe or whether it will lead to a dangerous contact with the environment. In light of this observation, control solutions are needed to facilitate safe and healthy locomotion over common and frequent barriers such as curbs or stairs. A notable challenge for intelligent prosthetics to overcome is therefore the ability sense and act upon important features in the environment.

[0005] Prior work in the field has centered on identifying discrete terrain classes based on kinematics including slopes, stairs, and uneven terrain. Vision systems in the form of depth sensors have recently been utilized in several vision-assisted exoskeleton robots. However, depth sensors with sufficient accuracy at close range are not portable, e.g., Li-DAR, and often prohibitively expensive. There is a current lack of solutions that provide high fidelity depth sensing and portability for use in environment-aware prosthetics.

[0006] It is with these observations in mind, among others, that various aspects of the present disclosure were conceived and developed.

BRIEF DESCRIPTION OF THE DRAWINGS

[0007] The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

[0008] FIG. **1** is a simplified diagram showing a system for environment-aware generation of control signals for a prosthetic or orthotic joint;

[0009] FIGS. **2**A and **2**B are simplified diagrams showing a framework for generating a control model for the system of FIG. **1**;

[0010] FIG. **2**C is a simplified diagram showing a depth and segmentation model for the framework of FIG. **2**A;

[0011] FIG. **2**D is a simplified diagram showing an ensemble Bayesian interaction primitive generation model for the framework of FIG. **2**A;

[0012] FIG. **3** is a simplified diagram showing a depth prediction neural network for the framework of FIG. **2**A;

[0013] FIGS. **4**A and **4**B show a series of images showing human walking on different ground surfaces;

[0014] FIGS. **5**A and **5**B show a series of images showing validation of the framework of FIG. **2**A;

[0015] FIG. **6** shows a series of images illustrating a predicted depth map with respect to a ground truth depth map by the framework of FIG. **2**A;

[0016] FIG. **7** is a graphical representation showing prediction of the ankle angle control trajectory for an entire phase of walking by the system of FIG. **1**;

[0017] FIGS. **8**A and **8**B show a series of images illustrating depth estimation with point cloud view;

[0018] FIG. **9** is an image showing a 3D point cloud for depth estimation of a subject stepping on a stair;

[0019] FIG. **10** is a graphical representation showing prediction of an ankle angle control trajectory for a single step; and

[0020] FIGS. **11**A-**11**F are a series of process flows showing a method that implements aspects of the framework of FIGS. **2**A-**2**D;

[0021] FIG. **12** is a simplified diagram showing an exemplary computing system for implementation of the system of FIG. **1**.

[0022] Corresponding reference characters indicate corresponding elements among the view of the drawings. The headings used in the figures do not limit the scope of the claims.

DETAILED DESCRIPTION

[0023] Various embodiments of an environment-aware prediction and control system and associated framework for human-robot symbiotic walking are disclosed herein. The system takes a single, monocular RGB image from a leg-mounted camera to generate important visual features of the current surroundings, including the depth of objects and the location of the foot. In turn, the system includes a data-driven controller that uses these features to generate adaptive and responsive actuation signals. The system employs a data-driven technique to extract critical perceptual information from low-cost sensors including a simple RGB camera and IMUs. To this end, a new, multimodal data set was collected for walking with the system on variable ground across 57 varied scenarios, e.g., roadways, curbs, gravel, etc. In turn, the data set can be used to train modules for environmental awareness and robot control. Once trained, the system can process incoming images and generate depth estimates and segmentations of the foot. Together with kinematic sensor modalities from the prosthesis, these visual features are then used to generate predictive control actions. To this end, the system builds upon ensemble Bayesian interaction primitives (enBIP), which have previously been used for accurate prediction in human biomechanics and locomotion. However, going beyond prior work on Interaction Primitives, the present system incorporates the perceptual features directly into a probabilistic model formulation to learn a state of the environment and generate predictive control signals. As a result of this data-driven training scheme, the prosthesis automatically adapts to variations in the ground for mobility-related actions such as lifting a leg to step up a small curb.

Environment-Aware Prediction and Control System

[0024] Referring to FIGS. **1**-**4**B, an environment-aware prediction and control system **100** (hereinafter, system **100**) integrates depth-based environmental terrain information into a holistic control model for human-robot symbiotic walking. The system **100** provides a prosthetic or orthotic joint **130** in communication with a computing device **120** and a camera **110** that collectively enable the prosthetic or orthotic joint **130** to take environmental surroundings into account when performing various actions. In one example embodiment shown in FIG. **1**, the prosthetic or orthotic joint **130** is a powered, electronically assisted prosthetic or orthotic that includes an ankle joint. The prosthetic or orthotic joint **130** can receive one or more control signals from the computing device **120** that dictate movement of various sub-components of the prosthetic or orthotic joint **130**. As such, the prosthetic or orthotic joint **130** can be configured to assist a wearer in performing various mobility-related tasks such as walking, stepping onto stairs and/or curbs, shifting weight, etc.

[0025] The computing device **120** receives image and/or video data from the camera **110** which captures information about various environmental surroundings and enables the computing device **120** to make informed decisions about the control signals applied to the prosthetic or orthotic joint **130**. The computing device **120** includes a processor in communication with a memory, the memory including instructions that enable the processor to implement a framework **200** that receives the image and/or video data from the camera **110** as the wearer uses the prosthetic or orthotic joint **130**, extracts a set of depth features from the image and/or video data that indicate perceived spatial depth information of a surrounding environment, and determines a control signal to be applied to the prosthetic or orthotic joint **130** based on the perceived depth features. Following determination of the control signal by the framework **200** implemented at the processor, the computing device **120** applies the control signal to the prosthetic or orthotic joint **130**. The present disclosure investigates the efficacy of the system **100** by evaluating how well the prosthetic or orthotic joint **130** performs on tasks such as stepping onto stairs or curbs aided by the framework **200** of the computing device **120**. In some embodiments, the camera **110** can be leg-mounted or mounted in a suitable location that enables the camera **110** to capture images of an environment that is in front of the prosthetic or orthotic joint **130**. To achieve environmental awareness in human-robot symbiotic walking of the system **100**, the following is achieved: (a) perform visual and kinematic data collection from an able-bodied subject, (b) augment the data set with segmented depth features from a trained depth estimation deep neural network, and (c) train a probabilistic model to synthesize control signals to be applied to the prosthetic or orthotic joint **130** given perceived depth features.

[0026] The framework **200** implemented at the computing device **120** of system **100** is depicted in FIGS. **2**A-**2**D. The framework **200** is organized into two main sections including: a depth and segmentation module **210** (FIG. **2**C) that extracts the set of depth features indicative of spatial depth information of a surrounding environment from an image captured by the camera **110** and performs a foot segmentation task on the image; and a control output module **220** (FIG. **2**D) that generates control signals for the computing device **120** to apply to the prosthetic or orthotic joint **130** based on the set of depth features extracted by depth and segmentation module **210** from the image captured by the camera **110**. The depth and segmentation module **210** includes a depth estimation network **212** defining a network architecture, with loss functions and temporal consistency constraints that enable the depth and segmentation module **210** to estimate a pixel level depth map from image data captured by the camera **110**, while ensuring low noise and high temporal consistency. In some embodiments, the image data captured by the camera **110** includes an RGB value for each respective pixel of a plurality of pixels of the image data, which the depth and segmentation module **210** uses to extract depth features of the surrounding environment. As humans naturally use depth perception to modulate their own movements during mobility-related tasks, the system **100** extends this ability to the prosthetic or orthotic joint **130** by using depth perception to modulate control inputs applied to the prosthetic or orthotic joint **130**.

[0027] Referring to FIG. **2**D, the control output module **220** uses ensemble Bayesian interaction primitives (enBIP) to generate environmentally-adaptive control outputs via inference within a latent space based on the extracted depth features from the depth and segmentation module **210**. As detailed below, enBIP is an extension of interaction primitives which have been utilized extensively for physical human-robot interaction (HRI) tasks including games of catch, handshakes with complex artificial muscle based humanoid robots, and optimal prosthesis control. A critical feature of enBIPs is their ability to develop learned models of that describe coupled spatial and temporal relationships between human and robot partners, paired with powerful nonlinear filtering, which is why enBIPs work well in human-robot collaboration tasks. The combination of both these neural network modules as well as the learned enBIP model of the framework **200** can be implemented within the system **100** using off-the-shelf components such as a mobile Jetson Xavier board evaluated below.

Depth Prediction Network

[0028] Referring to FIGS. **2**A-**2**C and **3**, the depth estimation network **212** for depth prediction as implemented by the depth and segmentation module **210** for navigating terrain features is described herein. The depth estimation network **212** uses a combination of convolutional and residual blocks in an autoencoder architecture (AE) to generate depth predictions from RGB image data captured by the camera **110**.

[0029] Network Architecture: The depth estimation network **212**, shown in FIG. **3**, is an AE network which utilizes residual learning, convolutional blocks, and skip connections to ultimately extract a final depth feature estimate f that describe depth features within an image I_t captured by the camera **110**. The depth estimation network **212** starts with a encoder network **214** (in particular, a ResNet-50 encoder network, which was shown to be a fast and accurate encoder model) and includes a decoder network **216**. Layers of the encoder network **214** and the decoder network **216** are connected via skip connection in a symmetrical manner to provide additional information at decoding time. Finally, the depth estimation network **212** uses a DispNet training structure to implement a loss weight schedule through downsampling. This approach enables the depth estimation network **212** to first learn a coarse representation of depth features from digital RGB images to constrain intermediate features during training, while finer resolutions impact the overall accuracy.

[0030] The depth feature extraction process starts with an input image I_t captured by the camera **110**, where $I \in \mathbb{R}^{H \times W \times 3}$ and where the image I_t includes RGB data for each pixel therewithin. The input image I_t is provided to the depth estimation network **212**. The encoder network **214** of the depth estimation network **212** receives the input image I_t first. As shown, in one embodiment, the encoder network **214** includes five stages. Following a typical AE network architecture, each stage of the encoder network **214** narrows the size of the representation from 2048 neurons down to 64 neurons at a final convolutional bottleneck layer.

[0031] The decoder network **216** increases the network size at each layer after the final convolutional bottleneck layer of the encoder network **214** in a pattern symmetrical with that of the encoder network **214**. While the first two stages of the decoder network **216** are transpose residual blocks of 3×3 size kernels, the third and fourth stages of the decoder network **216** are convolutional projection layers with two 1×1 kernels each. Although ReLU activation functions connect each stage of the decoder network **216**, additional sigmoid activation function outputs facilitate disparity estimation in the decoder network **216** by computing the loss for each output of varied resolutions. The output of the depth estimation network **212** is a combination of a final depth feature estimate $\hat{f} \in \mathbb{R}^{H \times W}$ for the input image I_t, as well as intermediate feature map outputs $\hat{f}_n \in \mathbb{R}^{HR_n \times WR_n}$ with $n \leq 5$ for each stage of the decoder network **216** starting at the final convolutional bottleneck layer of the encoder network and $R=[r^{-5}, r^{-4}, r^{-3}, r^{-2}, r^{-1}]$ being a resolution modifier defined by a resolution coefficient r. In one implementation shown in FIG. **3**, the decoder network of the depth estimation network **212** provides 1 output and 5 hidden feature map predictions in different resolutions, including estimated depth values for each pixel within the input image I_t, with the combination of all outputs denoted as $\hat{D}$ where $\hat{D}=[\hat{f}_5, \hat{f}_4, \hat{f}_3, \hat{f}_2, \hat{f}_1\ \hat{f}]$.

[0032] Loss Function: In order to use the full combination of final and intermediate outputs of D in a loss function of the depth estimation network **212** during training, it is necessary to first define a loss £ as a summation of a depth image reconstruction loss ER over each of the prediction of various resolutions (e.g., the final feature map output from a final output layer of the decoder network **216** in addition to intermediate feature map outputs). The loss £ can be described as:

$$\mathcal{L} = \sum_{i=1}^{6} \mathcal{L}_R(D_i \hat{D}_i) w'_i, \tag{1}$$

where estimated depth values at each stage D of the decoder network **216** are compared to a ground truth vector D, downsampled for each corresponding feature map resolution using average pooling operations. The depth estimation network **212** uses a loss weight vector to adjust for each feature size with associated elements of w'=[1/64, 1/32, 1/16, 1/8, 1/4, 1/2]. The depth estimation network **212** uses a reconstruction loss $\mathcal{L}_R$ for depth image reconstruction that includes four loss elements, including: (1) a mean squared error measure $\mathcal{L}_m$, (2) a structural similarity measure $\mathcal{L}_s$, (3) an inter-image gradient measure $\mathcal{L}_g$, and (4) a total variation measure $\mathcal{L}_{tv}$:

$$(d, \hat{d}) = \alpha_1 \mathcal{L}_m + \alpha_2 \mathcal{L}_s + \alpha_3 \mathcal{L}_g + \alpha_4 \mathcal{L}_{tv} \tag{2}$$

where $\hat{d}$ is an iterated depth feature output $\hat{d} \in \hat{D}$ and d is a ground truth feature $d \in D$, and hyperparameters $\alpha_1=102$, $\alpha_2=1$, $\alpha_3=1$, and $\alpha_4=10^{-7}$, influence the importance of each respective loss element on $\mathcal{L}_R$. The mean squared error measure can be represented as:

$$\mathcal{L}_m(d, \hat{d}) = \|d - \hat{d}\|_2^2 \tag{3}$$

[0033] A structural similarity index measure (SSIM) is adopted since it can be used to avoid distortions by capturing a covariance alongside an average of a ground truth feature map and a predicted depth feature map. A SSIM loss $\mathcal{L}_s$ can be represented as:

$$\mathcal{L}_s(d, \hat{d}) = \frac{1 - SSIM(d, \hat{d})}{2}. \tag{4}$$

[0034] Losses based on inter-image gradient primarily ensure illumination-invariance such that bright lights or shadows cast across the image do not affect the end depth prediction. The inter-image gradient measure $\mathcal{L}_g$ can be implemented as:

$$\mathcal{L}_g(d, \hat{d}) = \frac{1}{n} \sum_{i=1}^{n} (\|\nabla_x d_i - \nabla_x \hat{d}_i\| + \|\nabla_x d_i - \nabla_y \hat{d}\|, \tag{5}$$

where ∇ denotes a gradient calculation, and $\|\cdot\|$ is the absolute value. Since $\mathcal{L}_g$ is computed pixelwise both horizontally and vertically, a number of pixels of the output image is denoted as n. In order to add an additional loss to facilitate image deblurring and denoising, the total variation measure $\mathcal{L}_{tv}$ passes the gradient strictly over the output depth feature maps to minimize noise and round terrain features:

$$\mathcal{L}_{tv}(\hat{d}) = \left\|\nabla_x \hat{d}\right\|_1 + \left\|\nabla_y \hat{d}\right\|_1. \tag{6}$$

[0035] Temporal Consistency: Prediction consistency over time is a critical necessity for stable and accurate control of a robotic prosthesis. Temporal consistency of the depth predictions provided by framework **200** is achieved during training via the four loss functions, $\mathcal{L}^{m}$, $\mathcal{L}_{s}$, $\mathcal{L}_{g}$, and $\mathcal{L}_{tv}$ by fine-tuning the depth estimation network **212**. In particular, the framework **200** fine-tunes the depth estimation network **212** through application of a temporal consistency training methodology to the resultant depth feature output $\hat{d} \in \hat{D}$, which includes employing binary masks to outline one or more regions within an image which require higher accuracy, and further includes applying a disparity loss $\mathcal{L}_{dis}$ between two consecutive frames including a first frame taken at time t−1 and a second frame taken at time t (e.g., images within video data captured by camera **110**). An overlapping mask of the two frames can be defined as M and can be set to equal to the size of a ground truth feature d. As such, the disparity loss $\mathcal{L}_{dis}$ is formulated as:

$$\mathcal{L}_{dis}(\hat{d}_{t=1}, \hat{d}_t) = \beta_1 \mathcal{L}_m\left(\hat{d}_{t-1}^M, \hat{d}_t^M\right) + \beta_2 \mathcal{L}_s\left(\hat{d}_{t-1}^M, \hat{d}_t^M\right) + \gamma\left(\mathcal{L}_R(d_{t-1}, \hat{d}_{t-1}) + \mathcal{L}_R(d_t, \hat{d}_t)\right). \tag{7}$$

[0036] where $\hat{d}_t^M$ indicates for a predicted frame at time t with a binary mask applied: $\hat{d}_t^M = \hat{d}_t \cdot M$. Additionally, to sustain the precision of prediction over time, the framework **200** applies $\mathcal{L}_R$ is on each frame individually. Each loss element of the fine-tuning process is weighted by corresponding hyperparameters, $\beta_1=0.7$, $\beta_2=0.3$, $\gamma–10$. The fine-tuning step, therefore, makes the predicted frames more similar to one another in static regions while maintaining the reconstruction accuracy from prior network training.

Control Outputs Using Ensemble Bayesian Interaction Primitives

[0037] Given the extracted environmental information provided by depth and segmentation module **210** including depth features $\hat{D}=[\hat{f}_5, \hat{f}_4, \hat{f}_3, \hat{f}_2, \hat{f}_1\ \hat{f}]$. for each pixel within images captured by the camera **110**, the control output module **220** uses enBIP to generate appropriate responses for the prosthetic or orthotic joint **130**. As a data-driven method, enBIP uses example demonstrations of interactions between multiple agents to generate a behavior model that represents an observed system of human kinematics with respect to the prosthetic or orthotic joint **130**. enBIP was selected as a modeling formulation for this purpose because enBIP enables inference of future observable human-robot states as well as non-observable human-robot states. Additionally, enBIP supplies uncertainty estimates which can allow a controller such as computing device **120** to validate predicted control actions, possibly adding modifications if the model is highly unsure. Lastly, enBIP provides robustness against sensor noise as well as real-time inference capabilities in complex human-robot interactive control tasks.

[0038] In one aspect, assisted locomotion with a prosthetic or orthotic is cast as a close interaction between the human kinematics, environmental features, and robotic prosthetic. The control output module **220** incorporates environmental information in the form of predicted depth features along with sensed kinematic information from an inertial measurement unit (IMU) **160** (FIGS. **2**A and **2**B) and prosthesis control signals into a single holistic locomotion model. The control output module **220** uses kinematic sensor values, along with processed depth features and prosthesis control signals from n E N observed behavior demonstrations (e.g., demonstrated strides using the prosthetic or orthotic joint **130**), to form an observation vector $[y_1, \ldots y_{Tn}], \in \mathbb{R}^{D_y \times T_n}$ for D_y variables in T_n time steps. As such, by capturing the observed behavior demonstrations of the prosthetic or orthotic joint **130**, the control output module **220** can incorporate human kinematic properties and environmental features within the latent space to generate appropriate control signals that take into account human kinematic behavior especially in terms of observable environmental features (which can be captured at the depth and segmentation module **210**).

[0039] Latent Space Formulation: Generating an accurate model from an example demonstration matrix would be difficult due to high internal dimensionality, especially with no guarantee of temporal consistency between demonstrations. One main goal of a latent space formulation determined by control output module **220** is therefore to reduce modeling dimensionality by projecting training demonstrations (e.g., recorded demonstrations of human-prosthesis behavior) into a latent space that encompasses both spatial and temporal features. Notably, this process must be done in a way that allows for estimation of future state distributions for both observed and unobserved variables $Y_{t+1:T}$ with only a partial observation of the state space and the example demonstration matrix Y1/1:T_1, YN/1:T_N:

$$p\left(\hat{Y}_{t+1:T} \,\middle|\, y_{1:t}, Y^1_{1:T_1}, \ldots , Y^N_{1:T_N}\right). \tag{8}$$

[0040] Basis function decomposition sidesteps the significant modeling challenges of requiring a generative model over all variables and a nonlinear transition function. Basis function decomposition enables the control output module **220** to approximate each trajectory as a linear combination of B^d functions in the form of: $Y_t^d = \Phi_{\phi(t)}{}^T w^d + \in_y$. Each basis function $\Phi_{\phi(t)} \in \mathbb{R}^{B^d}$ is modified with corresponding weight parameters $w^d \in \mathbb{R}^{B^d}$ to minimize an approximation error $\in_y$. As a time-invariant dimensionality reduction method, the control output module **220** includes a temporal shift to the relative time measure phase $\phi(t) \in \mathbb{R}$, where $0 \le \phi(t) \le 1$.

[0041] Although the time-invariant latent space formulation facilitates estimation of entire trajectories, filtering over both spatial and temporal features interactions was more robust and accurate. As such, the control output module **220** incorporates phase, phase velocity, and weight vectors, $w=[\phi, \dot{\phi}, w^{0_T}, \ldots , w^{D_T}] \in \mathbb{R}^B$ where $B=\Sigma_d{}^D B^d$ into the state representation. By assuming that a training demonstration advances linearly in time, the phase velocity is estimated with, $\phi=1/T_n$. Substituting the weight vector into 8 and applying the Bayes' rule yields:

$$p(w_t Y_{1:t}, w_0) \propto p(y_t \,|\, w_t) p(w_t Y_{1:t}, w_0) \tag{9}$$

since the time-invariant weight vector p(w_t $Y_{1:t}$, w_0), models the entire trajectory Y.

[0042] Inference: In order to accommodate a variety of control modifications based on observed or predicted environmental features, the control output module **220** leverages ensemble Bayesian estimation from enBIP to produce approximate inferences of the posterior distribution according to Equation (9), which include human kinematics and environmental features. Assuming, of course, that higher-order statistical moments between states are negligible and that the Markov property holds. Algorithmically, enBIP first generates an ensemble of latent observation models, taken randomly from the demonstration set. As the subject walks with the prosthetic or orthotic joint **130**, the control output module **220** propagates the ensemble forward one step with a state transition function. Then, as new sensor and depth observations periodically become available as the camera **110** and the depth and segmentation module **210** work, the control output module **220** performs a measurement update step across the entire ensemble. From the updated ensemble, the control output module **220** calculates the mean and variance of each latent component, and subsequently projects the mean and variance into a trajectory space by applying the linear combination of basis functions to the weight vectors.

[0043] The control output module **220** uniformly samples the initial ensemble of E members X=[x^1, . . . , x^E] from the observed demonstrations x_0^j=[0, ϕ_i, w_i], 1≤j≤E with i~$\mathcal{U}$(1,N), and E≤N. Inference through Bayesian estimation begins as the control output module **220** iteratively propagates each ensemble member forward one step to approximate p(w_t|y_{t-1},w_0) with:

$$x_{t|t-1}^j = g\left(x_{t-1|t-1}^j\right) + \in_x,\ 1 \le j \le E, \tag{10}$$

which utilizes a constant-velocity state transition function g(·) and stochastic error $\in_x$≈N(0, Q_t), estimated with a normal distribution from the sample demonstrations. Next, the control output module **220** updates each ensemble member with the observation through the nonlinear observation operator h(·),

$$H_t X_{t|t-1} = \left[h\left(x_{t|t-1}^1\right), \dots , h\left(x_{t|t-1}^E\right)\right]^T. \tag{11}$$

followed by computing a deviation of the ensemble from the sample mean:

$$H_t A_t = H_t X_{t|t-1} - \left[\frac{1}{E}\sum_{j-1}^{E} h\left(x_{t|t-1}^j\right), \dots , \frac{1}{E}\sum_{j=1}^{E} h\left(x_{t|t-1}^j\right)\right]. \tag{12}$$

[0044] The control output module **220** uses the deviation H_tA_t and observation noise R to compute an innovation covariance:

$$w_t = \frac{1}{E-1}(H_t A_t)(H_t A_t)^T + R. \tag{13}$$

[0045] The control output module **220** uses the innovation covariance as well as the deviation of the ensemble to calculate the Kalman gain from the ensemble members without a covariance matrix through:

$$A_t = X_{t|t-1} - \frac{1}{E}\sum_{j=1}^{E} x_{t|t-1}^t, \tag{14}$$

$$K_t = \frac{1}{E-1} A_t (H_t A_t)^T w_t^{-1}. \tag{15}$$

[0046] Finally, the control output module **220** realizes a measurement update by applying a difference between a new observation at time t and an expected observation given t−1 to the ensemble through the Kalman gain,

$$\tilde{y}_t = \left[y_t + \epsilon_y^1, \dots , y_t + \epsilon_y^E\right], \tag{16}$$

$$X_{t|t} = X_{t|t-1} + K\left(\tilde{y}_t - H_t X_{t|t-1}\right). \tag{17}$$

[0047] The control output module **220** accommodates for partial observations by artificially inflating the observation noise for non-observable variables such as the control signals such that the Kalman filter does not condition on these unknown input values.

Experiments and Results

[0048] To validate the system **100**, a number of experiments were conducted with a focus on real-world human-subject data. The following describes in detail how the data was collected, processed, and how the models were trained, including specific hardware and software utilized. To better discuss specific model results, the following experimental section is further broken up into experiments and results for (1) the network architecture and (2) the enBIP model with environmental awareness. Experiment 1 examines the efficacy of our network architecture in predicting accurate depth values from RGB images collected from a body mounted camera. While Experiment 2 applies the network architecture on embedded hardware to infer environmentally conditioned control trajectories for a lower-limb prosthesis in a testbed environment.

Data Collection

[0049] Multimodal data sets were collected from participants who were outfitted with advanced inertial measurement units (IMUs) and a camera/depth sensor module. The IMUs are BNO080 system-in-package and include a triaxial accelerometer, a triaxial gyroscope, and a magnetometer with a 32-bit ARM Cortex microcontroller running Hillcrest Labs proprietary SH-2 firmware for sensor filtering and fusion. IMU devices are combined with an ESP32 microprocessor, in ergonomic cases that can easily be fitted to subjects' bodies over clothing, to send Bluetooth data packages out at 100 Hz. These inertial sensor modules were mounted to the subjects' lower limb and foot during the data collection process to collect kinematic data. However, during the testing phase, the foot sensor is removed. Additionally, an Intel RealSense D435 depth camera module was mounted to the subjects' lower limb.

[0050] A custom vision data set of 57 varied scenes was collected with over 30,000 RGB-depth image pairs from the lower-limb, during locomotion tasks in a dynamic urban environment. Data collection involved a subject walking over various obstacles and surfaces, including, but not limited to: sidewalks, roadways, curbs, gravel, carpeting, and up/down stairs; in differing lighting conditions at a fixed depth range (0.0-1.0 meters). Example images from the custom data set are visible in the upper row of FIGS. **4**A and **4**B with images of the subject walking in various scenes. The second row of FIGS. **4**A and **4**B visualizes the depth values of the same image using a color gradient; darker coloring indicates closer pixels while lighter pixels are further away. A custom annotated mask is shown in the third row of FIGS. **4**A and **4**B. These masks are used to train a masked RCNN model predicting human stepping areas. The semantic masks are also implemented to improve the temporal consistency.

Experiment 1: Neural Network Evaluation

[0051] In order for the network architecture of depth estimation network **212** to operate under real-world conditions, it must have both a low depth-prediction error, as well as real-time computation capabilities. The following section details the learning process and accuracy of our network architecture on the custom human-subject data set.

TABLE 1

Comparisons of different decoder architectures on the custom dataset.
Result Evaluations and Ablation Study

Encoder	Decoder	abs REL ↓	sq REL ↓	RMSE ↓	RMSE log ↓	δ_1↑	δ_2↑	δ_3↑	TC↓
ResNet-50	Residual Blocks	0.2292	0.0530	0.0841	0.0615	0.8144	0.9140	0.9575	0.000194
	Conv Blocks + DispNet	0.2285	0.0504	0.0808	0.0605	0.8057	0.9423	0.9656	0.000184
	Residual Blocks + Dispnet	0.2250	0.0494	0.0856	0.0633	0.7860	0.9175	0.9643	0.000195
	System 100	0.2205	0.0480	0.0753	0.0567	0.8227	0.9227	0.9658	0.000179

[0052] The last row shows results from the present depth estimation model. **1** indicates the higher the better; \ indicates the lower the better.

[0053] Training: 80% (46 scenes) of the custom data set was utilized for training and 20% (11 scenes) were utilized for testing. Adam was selected as the optimizer with learning rate $n_1=10^{-5}$. The input has the shape 90×160×3, whereas the ground truth and the output have the shape of 90×160×1. The ground truth is down-sampled to 3×5, 6×10, 12×20, 23×40, and 45×80 for the loss weight schedule of DispNet. Training was performed on 3 other AE architectures for comparison in an empirical manner. Residual learning, DispNet, and the combination of using convolutional layers are investigated for the decoder network **216** (see Table. I). All models were trained for 100 epochs and fine-tuned by applying a disparity loss with learning rate $n_2=10^{-7}$ for 30 epochs. FIGS. **5**A and **5**B show the validation among 4 models using the absolute REL and RMSE metrics.

[0054] A pre-trained masked RCNN was used as the masking network for object detection. The masked RCNN was fine-tuned given masks provided from the custom dataset using binary cross-entropy loss.

[0055] Results: The depth prediction results shown in FIG. **6** demonstrate prediction accuracy while walking in different conditions. Comparing the middle row and the bottom row of FIG. **6**, it can be seen that the predicted depth images exhibit a smoothing effect with less pixel to pixel noise than the original input. Additionally, sharp features in the environment, such as the stairs or curbs result in a delectably sharper edge between the two surfaces. Finally, the depth prediction network is shown to be accurate over a range of ground materials and patterns, such as solid bricks, concrete, aggregate gravel, carpet, and laminate stairs. The present depth-prediction network achieves depth estimation regardless of the ground material, lighting, or environmental features. It is particularly important to note that shadows (which are typically induced by the user) do not negatively affect the depth prediction performance.

[0056] Evaluation: The evaluation and the ablation study of the depth prediction results are shown in Table I, the evaluation process takes the RGB data from testing set and compared the model predictions with the ground truth in terms of commonly accepted evaluation metrics-absolute REL, sq REL, RMSE, RMSE log, δ_1, δ_2 and δ_3, where δ_N, is the percentage of the ground truth pixels under the constraint: $\max(\hat{d}_i/d_i, d_i/\hat{d}_i) < 1.25^N)$ since the for temporal consistency, $\mathcal{T}$C is proposed as the metric for consistency evaluation:

$$\mathcal{T}C = \frac{1}{T-1}\sum_{t=2}^{T}\left(\frac{1}{\Sigma_{i=1}^{n} M_i}\left\|\hat{d}_{t-1}^{M} - \hat{d}_{t}^{M}\right\|_2\right), \tag{18}$$

where T is the number of frames in a video sequence. Since the terrain from the camera view is dynamic, we conclude the lower $\mathcal{T}$C the better under the constrain: $\mathcal{T}C_{\approx}{}^{>}0$. Another visual way in evaluating depth estimation models is to review 3D point clouds generated with predicted depth maps. FIG. **8**B shows one sample point cloud for an image (FIG. **8**A) of the NYU-v2 data set (on which the system **100** did not train). In a similar vein, FIG. **9** depicts point clouds generated during an example task of stepping on a stair.

[0057] Because the final version of the framework **200** must integrate with the prosthetic or orthotic joint **130** and it must be capable of fast inference over a range of environmental variables. Therefore, the framework **200** was deployed on an embedded hardware serving as the computing device **120**, which is in some embodiments a Jetson Xavier NX, which is a system on module (SOM) device capable of up to 21 TOPS of accelerated computing and tailored toward streaming data from multiple sensors into modern deep neural networks. The framework **200** performed inference in an average of 0.0860 sec (11.57 FPS) with a standard deviation of 0.0013 sec.

Experiment 2: Prosthesis Evaluation

[0058] One critical use of environmental and terrain awareness is in stepping over curbs or onto stairs. If a terrain prediction algorithm is even 99% effective in stair prediction it would still pose a grave safety concern, due to the chances of causing the subject to fall down a set of stairs. Since environmental features are directly incorporated with a very high accuracy, the evaluation focuses experimentally on two criteria in the prosthesis experiments (1) "Can enBIP model stair stepping over a range of step distances?" and (2) "For a given step, does incorporating environmental features produce a more accurate model?".

[0059] Training: Collected data for stair stepping was used to train an enBIP model with modalities from tibia-mounted inertial sensors, predicted depth features, and the ankle angle control trajectory. To produce depth features from the predicted depth map, the system **100** took the average over two masks which bisect the image horizontally and subtracting the area behind the shoe from the area in front. A one-dimensional depth feature was produced which showed the changes in terrain due to slopes or steps. While the depth features for this experiment were simplified, other and more complex features were possible, such as, calculating curvature, detecting stair edges, or incorporating the entire predicted depth map. Subjects were asked to perform 50 steps of stair-stepping up onto a custom built curb during which the subject was instructed to start from with their toe at varying positions away from the curb in a range from 0 inches to 16 inches. Applying the framework **200**, the system **100** ends up with a generic model to predict ankle angle control actions given IMU observations and depth features. The compiled point cloud in FIG. **9** from one example demonstration illustrates the accuracy of the depth prediction method of depth and segmentation module **210** during experimentation.

[0060] Results: The system **100** produced an average control error of 1.07° over 10 withheld example demonstration when using depth features for the stair stepping task compared to an average control error of 6.60° without depth features. The system **100** performed even better when examined at 35% phase, the approximate temporal location where the foot traverses the leading edge of the stair, with average control error of 2.32° compared to 9.25° for inference with kinematics only. FIG. **10** highlights the difference between inference with and without environmental awareness, where from the partial observation of a stepping trajectory the system **100** produced two estimates of the control trajectory both with and without environmental awareness. Inference with environmental awareness (blue) has a low variance, which shows high confidence by the model, and withdraws the toe substantially based on the observed depth features. However, without environmental awareness (green) the model does not have adequate information to form an inference with high confidence leading to both an increase in variance, as well as a control trajectory which does not sufficient dorsiflexion to the foot to clear the curb.

[0061] FIG. **7** shows the range of possible control actions at the start of the stride given the position of the step as seen through the predicted depth features. Given the same initial conditions the horizontal position of the step in reference to the foot clearly modifies the ankle angle control trajectory. When the step is very close, the ankle angle must increase dramatically to pull the toe up such that there is no collision with the edge of the step. Likewise, as the step is moved away from the foot the ankle must first apply plantarflexion to push off the ground before returning to the neutral position before heel-strike.

Methods

[0062] FIGS. **11**A-**11**F are a series of process flows showing a method **300** for implementing aspects of the system **100** and associated framework.

[0063] At block **302** of method **300** shown in FIG. **11**A, a processor receives image data from a camera associated with a prosthetic or orthotic joint. At block **304**, the processor extracts, by a depth estimation network formulated at the processor, a set of depth features indicative of perceived spatial depth information of a surrounding environment from the image data. At block **306**, the processor generates, by a control output module formulated at the processor, a control signal to be applied to the prosthetic or orthotic joint by inference within a latent space based on the set of depth features, an orientation of the prosthetic or orthotic joint and an observed behavior model using ensemble Bayesian interaction primitives. At block **308**, the processor applies the control signal to the prosthetic or orthotic joint.

[0064] With reference to FIG. **11**B, block **304** includes a sub-block **340**, at which the processor estimates, by the depth estimation network, a pixel level depth map from the image data captured by the camera. Sub-block **340** includes sub-block **341** in which the processor extracts, by an encoder network of a depth estimation network formulated at the processor, an initial representation indicative of depth information within the image data. At sub-block **342**, the processor reconstructs, by a decoder network of the depth estimation network, the image data using the initial representation indicative of depth information within the image data. At sub-block **343**, the processor generates, by the depth estimation network, a predicted depth feature map including a set of depth features indicative of perceived spatial depth information of a surrounding environment from the image data. At block **344**, the processor segments, by a segmentation module in communication with the depth estimation network, the pixel level depth map into a first area and a second area, the first area including a limb associated with the prosthetic or orthotic joint and the second area including an environment around the limb.

[0065] With reference to FIG. **11**C, block **306** includes a sub-block **361**, at which the processor determines, at the control output module, an observed behavior model based on one or more depth features of the pixel level depth map and an orientation of the prosthetic or orthotic joint. Sub-block **361** includes a further sub-block **362** at which the processor samples, at the control output module, an ensemble of latent observations from one or more observed behavior demonstrations of the prosthetic or orthotic joint that incorporate human kinematic properties and environmental features within the latent space, a trajectory of the prosthetic or orthotic joint being collectively described by a plurality of basis functions.

[0066] Continuing with FIG. **11**D, sub-block **363** includes iteratively propagating, at the control output module, the ensemble forward by one step using a state transition function as the prosthetic or orthotic joint operates. Sub-block **364** includes iteratively updating, at the control output module, one or more measurements of the ensemble using the set of depth features from the pixel depth map and an orientation of the prosthetic or orthotic joint. Sub-block **365** includes iteratively projecting, at the control output module, a mean and variance of one or more latent components of the ensemble into a trajectory space through the plurality of basis functions and based on the one or more measurements of the ensemble. Sub-block **366** includes updating, at the control output module, the control signal based on a difference between a new observation at a first time t and an expected observation at a second time t−1, the expected observation being indicative of one or more measurements of the ensemble taken at time t−1 and the new observation being indicative of one or more measurements of the ensemble taken at time t. This control signal can be applied to the prosthetic or orthotic joint at block **308** of FIG. **11**A.

[0067] FIG. **11**E shows training the depth estimation network using a ground truth dataset, which is outlined at block **310**. At sub-block **311**, the processor applies the depth estimation network to the ground truth dataset that includes depth information for each of a plurality of images. At sub-block **312**, the processor determines, a loss associated with a decoder stage of a plurality of stages of the decoder network and an associated encoder stage of the depth estimation network. At sub-block **313**, the processor minimizes a loss between a ground truth feature and a depth feature of the set of depth features. At block **314**, the processor updates one or more parameters of the depth estimation network based on the loss between the ground truth feature and the depth feature of the set of depth features.

[0068] Continuing with FIG. **11**F, at sub-block **315**, the processor applies a temporal consistency training methodology to the set of depth features. Sub-block **315** can be divided further into sub-blocks **316**, **317** and **318**. At sub-block **316**, the processor outlines one or more regions in the image data that require higher accuracy. At sub-block **317**, the processor determines a disparity loss between a first frame of the image data taken at time t and a second frame of the image data taken at t−1. At sub-block **318**, the processor updates one or more parameters of the depth estimation network based on the disparity loss between the first frame and the second frame.

Computer-Implemented System

[0069] FIG. **12** is a schematic block diagram of an example computing device **400** that may be used with one or more embodiments described herein, e.g., as a component of system **100** and/or implementing aspects of framework **200** in FIGS. **2**A-**2**D and/or method **300** in FIGS. **11**A-**11**F.

[0070] Device **400** comprises one or more network interfaces **410** (e.g., wired, wireless, PLC, etc.), at least one processor **420**, and a memory **440** interconnected by a system bus **450**, as well as a power supply **460** (e.g., battery, plug-in, etc.).

[0071] Network interface(s) **410** include the mechanical, electrical, and signaling circuitry for communicating data over the communication links coupled to a communication network. Network interfaces **410** are configured to transmit and/or receive data using a variety of different communication protocols. As illustrated, the box representing network interfaces **410** is shown for simplicity, and it is appreciated that such interfaces may represent different types of network connections such as wireless and wired (physical) connections. Network interfaces **410** are shown separately from power supply **460**, however it is appreciated that the interfaces that support PLC protocols may communicate through power supply **460** and/or may be an integral component coupled to power supply **460**.

[0072] Memory **440** includes a plurality of storage locations that are addressable by processor **420** and network interfaces **410** for storing software programs and data structures associated with the embodiments described herein. In some embodiments, device **400** may have limited memory or no memory (e.g., no memory for storage other than for programs/processes operating on the device and associated caches).

[0073] Processor **420** comprises hardware elements or logic adapted to execute the software programs (e.g., instructions) and manipulate data structures **445**. An operating system **442**, portions of which are typically resident in memory **440** and executed by the processor, functionally organizes device **400** by, inter alia, invoking operations in support of software processes and/or services executing on the device. These software processes and/or services may include prosthetic or orthotic joint processes/services **490** described herein. Note that while prosthetic or orthotic joint processes/services **490** is illustrated in centralized memory **440**, alternative embodiments provide for the process to be operated within the network interfaces **410**, such as a component of a MAC layer, and/or as part of a distributed computing network environment.

[0074] It will be apparent to those skilled in the art that other processor and memory types, including various computer-readable media, may be used to store and execute program instructions pertaining to the techniques described herein. Also, while the description illustrates various processes, it is expressly contemplated that various processes may be embodied as modules or engines configured to operate in accordance with the techniques herein (e.g., according to the functionality of a similar process). In this context, the term module and engine may be interchangeable. In general, the term module or engine refers to model or an organization of interrelated software components/functions. Further, while the prosthetic or orthotic joint processes/services **490** is shown as a standalone process, those skilled in the art will appreciate that this process may be executed as a routine or module within other processes.

[0075] It should be understood from the foregoing that, while particular embodiments have been illustrated and described, various modifications can be made thereto without departing from the spirit and scope of the invention as will be apparent to those skilled in the art. Such changes and modifications are within the scope and teachings of this invention as defined in the claims appended hereto.

1. A system, comprising:

a prosthetic or orthotic joint configured to receive one or more control signals and operate in response to the one or more control signals;

a camera that captures image data indicative of a surrounding environment around the prosthetic or orthotic joint; and

a computing device in operative communication with the prosthetic or orthotic joint and the camera, the computing device including a processor in communication with a memory, the memory including instructions, which, when executed, cause the processor to:

receive, at the processor, the image data from the camera;

extract, by a depth estimation network formulated at the processor, a set of depth features indicative of perceived spatial depth information of a surrounding environment from the image data; and

generate, by a control output module formulated at the processor, a control signal to be applied to the prosthetic or orthotic joint based on the set of depth features.

2. The system of claim **1**, wherein the memory further includes instructions, which, when executed, cause the processor to:

estimate, by the depth estimation network, a pixel level depth map from the image data captured by the camera.

3. The system of claim **2**, wherein the memory further includes instructions, which, when executed, cause the processor to:

segment, by a segmentation module formulated at the processor and in communication with the depth estimation network, the pixel level depth map into a first area and a second area, the first area including a limb associated with the prosthetic or orthotic joint and the second area including an environment around the limb.

4. The system of claim **2**, wherein the memory further includes instructions, which, when executed, cause the processor to:

determine, at the control output module, the control signal for the prosthetic or orthotic joint based on the pixel level depth map and an observed behavior model.

5. The system of claim **4**, wherein the memory further includes instructions, which, when executed, cause the processor to:

determine, at the control output module, the observed behavior model based on one or more depth features of the pixel level depth map and an orientation of the prosthetic or orthotic joint.

6. The system of claim **5**, further comprising:

an inertial measurement unit in communication with the control output module that determines the orientation of the prosthetic or orthotic joint.

7. The system of claim **1**, wherein the memory further includes instructions, which, when executed, cause the processor to:

generate, at the control output module, one or more control signals by inference within a latent space based on the set of depth features and using ensemble Bayesian interaction primitives.

8. The system of claim **7**, wherein the memory further includes instructions, which, when executed, cause the processor to:

uniformly sample, at the control output module, an ensemble of latent observations from one or more observed behavior demonstrations of the prosthetic or orthotic joint that incorporate human kinematic properties and environmental features within the latent space, a trajectory of the prosthetic or orthotic joint being collectively described by a plurality of basis functions;

iteratively propagate, at the control output module, the ensemble forward by one step using a state transition function as the prosthetic or orthotic joint operates;

iteratively update, at the control output module, one or more measurements of the ensemble using the set of depth features and an orientation of the prosthetic or orthotic joint;

iteratively project, at the control output module, a mean and variance of one or more latent components of the ensemble into a trajectory space through the plurality of basis functions and based on the one or more measurements of the ensemble; and

update, at the control output module, the control signal based on a difference between a new observation at a first time t and an expected observation at a second time t−1, the expected observation being indicative of one or more measurements of the ensemble taken at time t−1 and the new observation being indicative of one or more measurements of the ensemble taken at time t.

9. The system of claim **1**, wherein the image data captured by the camera includes RGB image data, wherein each pixel of a plurality of pixels within the image data includes a corresponding RGB value.

10. A system, comprising:

a computing device including a processor in communication with a memory, the memory including instructions, which, when executed, cause the processor to:

receive, at the processor, image data from an image capture device;

extract, by an encoder network of a depth estimation network formulated at the processor, an initial representation indicative of depth information within the image data;

reconstruct, by a decoder network of the depth estimation network formulated at the processor, the image data using the initial representation indicative of depth information within the image data; and

generate, by the depth estimation network, a predicted depth feature map including a set of depth features indicative of perceived spatial depth information of a surrounding environment from the image data.

11. The system of claim **10**, wherein the image data includes a plurality of pixels and wherein the set of depth features include a depth prediction for one or more pixels of the plurality of pixels of the image data.

12. The system of claim **10**, wherein the depth estimation network is an autoencoder network.

13. The system of claim **10**, the decoder network including:

a plurality of decoder stages that result in the set of depth features based on the image data at varying resolutions between each respective decoder stage of the plurality of decoder stages, the plurality of decoder stages including at least one transpose residual block, at least one convolutional projection layer, and an output layer, each respective decoder stage of the plurality of decoder stages being associated with a respective encoder stage of a plurality of encoder stages of the encoder network.

14. The system of claim **13**, wherein the memory further includes instructions, which, when executed, cause the processor to:

determine, at the processor, a loss associated with a decoder stage of the plurality of decoder stages of the decoder network and an associated encoder stage of the encoder network.

15. The system of claim **10**, wherein the memory further includes instructions, which, when executed, cause the processor to:

minimize, by the processor, a loss between a ground truth feature and a depth feature of the set of depth features; and

update, by the processor, one or more parameters of the depth estimation network based on the loss between the ground truth feature and the depth feature of the set of depth features.

16. The system of claim **15**, the loss including:

a mean squared error measure indicative of an error between the ground truth feature and the depth feature of the set of depth features;

a structural similarity index measure indicative of a covariance and/or an average of a ground truth feature map and a predicted depth feature map indicative of the set of depth features;

an inter-image gradient measure indicative of a gradient difference between the ground truth feature map and the predicted depth feature map; and

a total variation measure indicative of a total variation ground truth feature map and the predicted depth feature map.

17. The system of claim **10**, wherein the memory further includes instructions, which, when executed, cause the processor to:

apply, at the processor, a temporal consistency methodology to the set of depth features.

18. The system of claim **17**, wherein the memory further includes instructions, which, when executed, cause the processor to:

outline one or more regions in the image data that require higher accuracy; and

determine a disparity loss between a first frame of the image data taken at time t and a second frame of the image data taken at t−1.

19. The system of claim **18**, wherein the memory further includes instructions, which, when executed, cause the processor to:

update, by the processor, one or more parameters of the depth estimation network based on the disparity loss between the first frame and the second frame.

20. A method, comprising:

receiving, at a processor, image data from a camera associated with a prosthetic or orthotic joint;

extracting, by a depth estimation network formulated at the processor, a set of depth features indicative of perceived spatial depth information of a surrounding environment from the image data;

generating, by a control output module formulated at the processor, a control signal to be applied to the prosthetic or orthotic joint by inference within a latent space based on the set of depth features, an orientation of the prosthetic or orthotic joint and an observed behavior model using ensemble Bayesian interaction primitives; and

applying, by the processor, the control signal to the prosthetic or orthotic joint.

* * * * *