Multimodal Technologies and Interaction

MDPI

Article

AI Technologies for Machine Supervision and Help in a Rehabilitation Scenario

Gábor Baranyi [1], Bruno Carlos Dos Santos Melício [1], Zsófia Gaál [1,2], Levente Hajder [1], András Simonyi [1], Dániel Sindely [1], Joul Skaf [1], Ondřej Dušek [3], Tomáš Nekvinda [3] and András Lőrincz [1,*]

[1] Department of Artificial Intelligence, Faculty of Informatics, Eötvös Loránd University, 1083 Budapest, Hungary; bagtabi@inf.elte.hu (G.B.); a1w636@inf.elte.hu (B.C.D.S.M.); zsofia.gaal@gmail.com (Z.G.); hajder@inf.elte.hu (L.H.); simonyi@inf.elte.hu (A.S.); sindely.daniel@gmail.com (D.S.); slnj33@inf.elte.hu (J.S.)

[2] Emineo Private Hospital, 1016 Budapest, Hungary

[3] Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, 118 00 Prague, Czech Republic; odusek@ufal.mff.cuni.cz (O.D.); nekvinda@ufal.mff.cuni.cz (T.N.)

* Correspondence: lorincz@inf.elte.hu

Abstract: We consider, evaluate, and develop methods for home rehabilitation scenarios. We show the required modules for this scenario. Due to the large number of modules, the framework falls into the category of Composite AI. Our work is based on collected videos with high-quality execution and samples of typical errors. They are augmented by sample dialogues about the exercise to be executed and the assumed errors. We study and discuss body pose estimation technology, dialogue systems of different kinds and the emerging constraints of verbal communication. We demonstrate that the optimization of the camera and the body pose allows high-precision recording and requires the following components: (1) optimization needs a 3D representation of the environment, (2) a navigation dialogue to guide the patient to the optimal pose, (3) semantic and instance maps are necessary for verbal instructions about the navigation. We put forth different communication methods, from video-based presentation to chit-chat-like dialogues through rule-based methods. We discuss the methods for different aspects of the challenges that can improve the performance of the individual components. Due to the emerging solutions, we claim that the range of applications will drastically grow in the very near future.

Keywords: body pose estimation; dialogue system; semantic map; instance map; real-time interaction; deep learning; Composite AI

check for updates

Citation: Baranyi, G.; Dos Santos Melício, B.C.; Gaál, Z.; Hajder, L.; Simonyi, A.; Sindely, D.; Skaf, J.; Dušek, O.; Nekvinda, T.; Lőrincz, A. AI Technologies for Machine Supervision and Help in a Rehabilitation Scenario. *Multimodal Technol. Interact.* **2022**, *6*, 48. https://doi.org/10.3390/mti6070048

Academic Editor: Christos Troussas, Cleo Sgouropoulou, Akrivi Krouska, Ioannis Voyiatzis and Athanasios Voulodimos

Received: 23 May 2022

Accepted: 16 June 2022

Published: 22 June 2022

1. Introduction

Human–machine, human–robot interactions are complex and need diverse computations, for example in the relatively simple case of home rehabilitation scenarios. We consider, evaluate, develop methods, and list the required modules for this scenario. We suggest combining them by means of Composite AI [1], i.e., we consider deep neural networks as sensors and the problem of merging their outputs using higher level knowledge. Tasks include the optimization of the camera position and the pose of the patient, a spatial task and the communication about the navigation and the exercises including some notes on potential errors. Communication in the simplest case can use the keyboard and the screen, e.g., for showing the exercises and questionnaires about performance, errors, and pains. More sophisticated solutions involve natural language processing (NLP) and speech. We assume that high-quality automated speech recognition and text-to-speech engines are available and consider image processing and NLP.

The utilization and market potential of physical rehabilitation applications is growing [2]. Rehabilitation and health care robotics have been developed over the years [3] and home-based mechatronic technologies have been proposed [4]. Mechatronic components constrain the spatial tasks, e.g., there is no need for navigation. They alleviate body motion detection too, due to the direct contact between the device and the body. Here, we aim to show that components for machine-assisted free-space exercises are within reach.

In our concrete rehabilitation scenario, we worked with the Emineo Private Hospital in Budapest who perform Total Knee and Hip Replacement operations and have physical therapists to help and guide the healing process. We collected 50 knee rehabilitation exercises and their potential errors at Emineo. This database includes videos and textual data.

The interaction with the system represents a physiotherapy session, similar to one a patient would have with a human physiotherapist. In the current version, we assume that the system aims to guide the patient through an exercise plan/schedule specified by a physiotherapist. The system keeps a profile for each user with the exercise history.

In what follows, we first turn to the spatial tasks to be followed by physical rehabilitation as an application.

1.1. General Spatial Tasks

Robots are playing an increasingly significant role in our everyday lives. As technology develops, robotic agents are starting to replace humans, among others, in assistive roles ranging from guiding visitors in large, unfamiliar buildings [5] to supporting elderly people in their home [6], or mitigating workforce shortages in health care [7]. For a systematic overview of the usage of robots in human care, see [8].

Providing effective support for human physical activities requires three important capabilities:

1. *Collaborative planning:* The assisting robot has to be able to plan the collaboration in advance, which, in turn, presupposes the ability to reason about the physical environment and the human to be assisted.
2. *Accurate perception:* The robot must be able to sense the environment and the movements of the human partner to be assisted, and suggest corrections with high accuracy when needed.
3. *Communication* is another necessary component. It may concern the routes to be taken, help in executing and optimizing the activities, error detection, among others. In the case of physical rehabilitation, the patient's data are to be collected. A questionnaire may be sufficient for this. The exercise and the patient's position should be communicated, e.g., by texting, by video-based instructions, by speech, or by a combination of these. Detected errors in the exercise can also be communicated. Apart from a limited video-based communication, both texting and speech call for Natural Language Processing, Understanding, and Generation.

In this paper, we put forth a framework composed of the technology elements needed to achieve these tasks and show how it can be used in a concrete rehabilitation scenario.

We call the framework of the collection of our modules STF, a shorthand for *Spatial Task Framework*. It is constructed for human–machine collaboration. STF modules can support fast and accurate collaboration for navigation, human motion planning, monitoring, evaluation, and error correction of different scenarios.

STF includes modules that deal with (a) the human in the environment, (b) the configuration of the body or the body parts, and (c) related communication tasks. (a) includes the learning of the environment and assisting in navigation. Item (b) deals with the perception of body position and its comparison with the prescribed body configuration, and may deal with body-object processes such as manipulation, which are not discussed here. Finally, (c) covers generating verbal navigation and movement instructions to the user.

The number of components is relatively large due to several factors that one has to deal with: (i) interaction is multi-modal, (ii) spatial features include position and pose for the objects and the human partner, (iii) verbal communication requires semantic and instance maps, and finally, (iv) not all technology components are sufficient for real-time interaction as of today. We consider items 1 and 2 from the numbered list above and separate the tasks that can be executed *offline* and may be slower from those that must run in *real time*. We study in detail the components that require both high accuracy and real-time execution.

Due to the large number of modules and their optimization to achieve the best results, the framework falls under the category of Composite AI [1].

1.2. Physical Rehabilitation as an STF Application

Rehabilitation has its own organizational, financial, and technological challenges [9], which make machine-assisted rehabilitation desirable. While rehabilitation or physical therapy is required in treatment for many different conditions as well as in post-operation recovery, rehabilitation sessions with the assistance of a physiotherapist are expensive, and may be time-consuming or even infeasible due to physical distance between the patient and the therapist. However, the success of a treatment, e.g., the success of a Total Knee or Hip Replacement (TKR and THR, respectively) operation and post-operative therapy depends strongly on the rehabilitation quality, continuous help, and personalization. Providing assistance during a rehabilitation session typically requires multimodal interaction, including the detection of errors, the estimation of the level of pain, expert advice on error correction, modifications of the planned rehabilitation trajectory, and explanatory dialogues for specific errors or pain conditions. Automating such assistance thus presents serious challenges.

1.3. Related Works

We review indoor navigation, 3D body pose estimation, and works related to physical rehabilitation.

1.3.1. Indoor Navigation Instruction Generation

Outdoor routing instruction generation has been the subject of intensive research since the appearance of location-based services. However, it was only relatively recently that generating indoor navigation instructions has also started to receive significant attention. One of the early milestones was the introduction of the GIVE (Generating Instructions in Virtual Environments) series of shared tasks [10]. The field is very active. Challenges include the 'PointGoal' and 'ObjectGoal' tasks, i.e., to navigate to a point or to an object as defined in [11]. A different, more complex task is the Watch-and-Help (WAH) task [12]. In this task, a robot agent observes a task's execution, infers the goal and in another environment helps with carrying out the same task, demonstrating generalization capabilities.

Our studies are less ambitious than the the aforementioned tasks: we consider a navigation scenario starting with a human request for help. The system responds with verbal machine instructions guiding the person to move towards the goal position using 3D dense representations of the space like semantic and instance maps or, in the case of method development, a simulation like Habitat: the semantic and instance maps enable the construction of verbal instructions. During navigation, the software selects distances, directions and salient features for verbal communication. The method has promise for real-life scenarios since similar dense semantic models of the space can be learned, see, e.g., [13] and the references therein.

1.3.2. 3D Body Pose and Shape

The estimation of 3D human pose from 2D images has made significant progress in the last few years. Algorithms approximate 2D landmarks, for example using OpenPose [14] or HRNet [15], followed by the estimation of 3D pose using deep networks (for an overview, see [16]).

MeTRAbs uses metric-scale truncation-robust (MeTRo) volumetric heatmaps, whose dimensions are all defined in metric 3D space allowing the direct estimation of complete, metric-scale poses. MeTRAbs is robust as it estimates spine joints beyond 24 joints. This deep backbone network is, however, slow for real-time application.

The Blazepose component of MediaPipe [17] estimations works in real time. It uses a lightweight convolutional neural network architecture. It outputs 33 body keypoint estimations and the visibility scores serving occlusion detection.

MeTRAbs and MediaPipe, like other skeleton-based methods, neglect important aspects of the body such as the shape that may introduce errors in depth because the location of joints can differ according to the body types. Efforts for high-quality animations brought about approaches that estimate human pose and shape together. Popular backbones are the SMPL [18] and SMPL-X [19] models. SMPL and SMPL-X are realistic 3D models of the human body based on skinning and blend shapes learned from thousands of 3D body scans. They accurately represent a wide variety of body shapes of natural human poses. They can also regress joint locations from their meshes.

SMPL-X facilitates the analysis of human actions, interactions, and emotions by approximating 3D models of the human body pose, hand pose, and facial expression from a monocular image. ExPose [20] regresses SMPL-X parameters by estimating 2D features and then optimizing model parameters to fit those features.

We combine model-free methods, such as MeTRAbs [21] or MediaPipe [22], with a Skinned Multi-Person Linear (SMPL) model, ExPose [20].

1.3.3. Assisting Physical Rehabilitation

Recent medical assistive technologies have made a great progress, especially with respect to detecting various conditions from video data and real-time musculoskeletal health monitoring (for a recent review, see [23] and the cited references therein). In addition, rehabilitation after TKR in clinician monitored home-based programs seems as good as inpatient rehabilitation [24], raising the desire for remotely monitored and machine assisted rehabilitation.

There have been multiple works specifically aimed at assessing therapy exercises and providing feedback [25–27]. However, the current approaches can only provide automatic motion detection or assessment based on wearable devices or video input. This is nowhere near as good as personal physiotherapy sessions, where the correct sequence of exercise is demonstrated, performance errors are noted as they occur and both verbal and visual corrections of movement are facilitated.

1.4. Our Framework

Figure 1 depicts the key elements of our Spatial Task Framework. We put forth a conceptual modular spatial task framework that can support spatial collaboration between human and machine partners.

Our contributions are as follows:

1. Verbal communication about the environment requires semantic and instance maps that can work in real time, and shared object semantics between the human and the guiding system. We review the components developed previously in [13].
2. We present a navigation method in a realistic home environment on the open source virtual platform Habitat 2.0 [28] including path planning, the optimization of the camera placement and starting body position and pose, and the NLP solution helping human navigation to the optimal position,
3. For providing feedback about the motion executed, we explore three options: (1) no dialogue (a video-only system), (2) a rule-based dialogue system based on expert advice, with a data-driven user intent detection mechanism, and (3) a fully data-driven dialogue system based on DialoGPT [29].
4. We collected data resources on 50 knee rehabilitation exercises and their potential errors. The database includes video and text materials and forms the basis of our methods.
5. An additional crowdsourced dataset of in-domain dialogues grounded in video examples was also collected and used.
6. We evaluate and discuss the strengths and weaknesses of both the video detection methods as well as all variants of the feedback system.

7. We present results for the following components: absolute 3D human pose skeleton estimation using (i) MeTRAbs [21] and (ii) MediaPipe [22], (iii) ExPose [20] that directly regresses the body, face, and hands, in SMPL-X format for the estimation of body shape, (iv) the point cloud of the Zed2 or Kinect Azure camera to derive ground truth values, and (v) a perspective distortion compensation module for the case if the camera and the body are too close to each other.

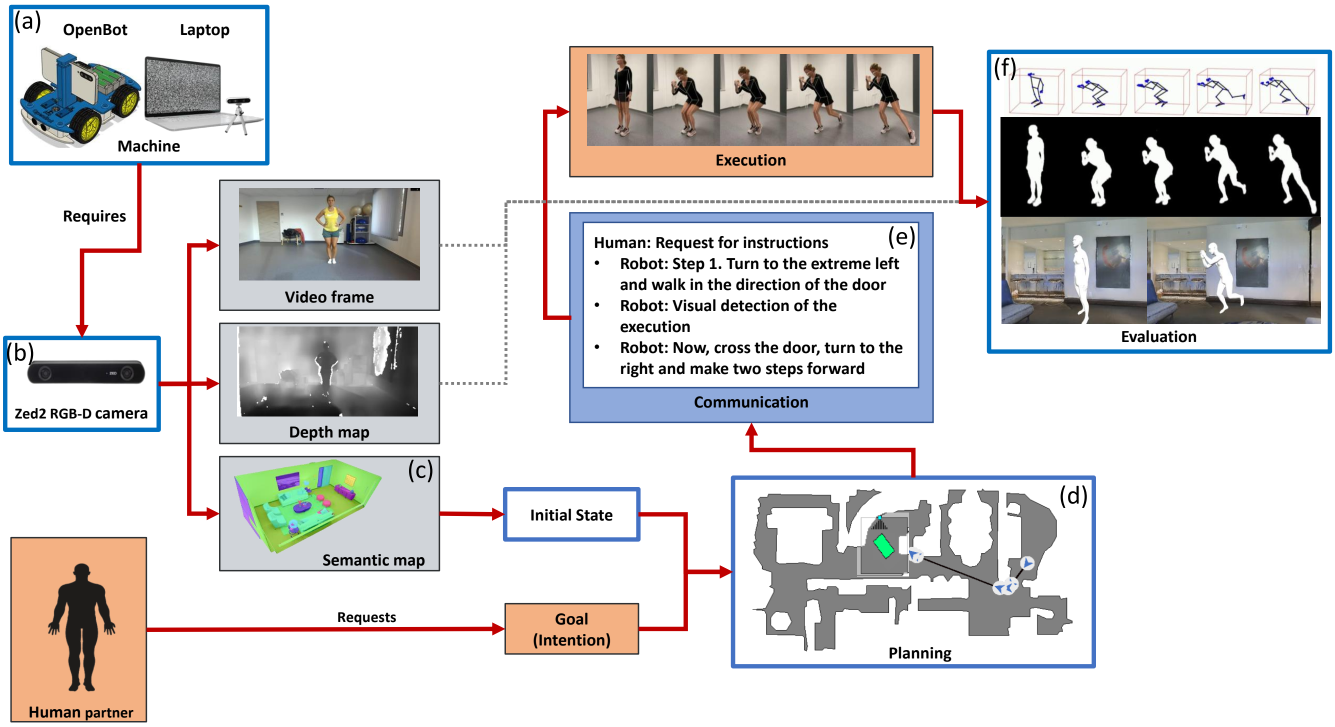

(a)
OpenBot
Laptop
Machine
Requires
(b)
Zed2 RGB-D camera
Video frame
Depth map
(c)
Semantic map
Human partner
Requests
Initial State
Goal (Intention)
Execution
(e)
Human: Request for instructions
• Robot: Step 1. Turn to the extreme left and walk in the direction of the door
• Robot: Visual detection of the execution
• Robot: Now, cross the door, turn to the right and make two steps forward
Communication
(d)
Planning
(f)
Evaluation

Figure 1. The Spatial Task Framework (STF) framework. Machine modules: (**a**) laptop, 3D-printed OpenBot robot [30], (**b**) Zed2 stereo or Kinect Azure ToF camera mounted on the robot base, (**c**) Simulation tool: Replica environments in Habitat platform [28], (**d**) a SLAM module for navigation planning, (**e**) a communication module for machine help in navigation, and (**f**) different body pose estimation methods, including an avatar for position optimization, body motion detection and evaluation (see Figure 2 for details).

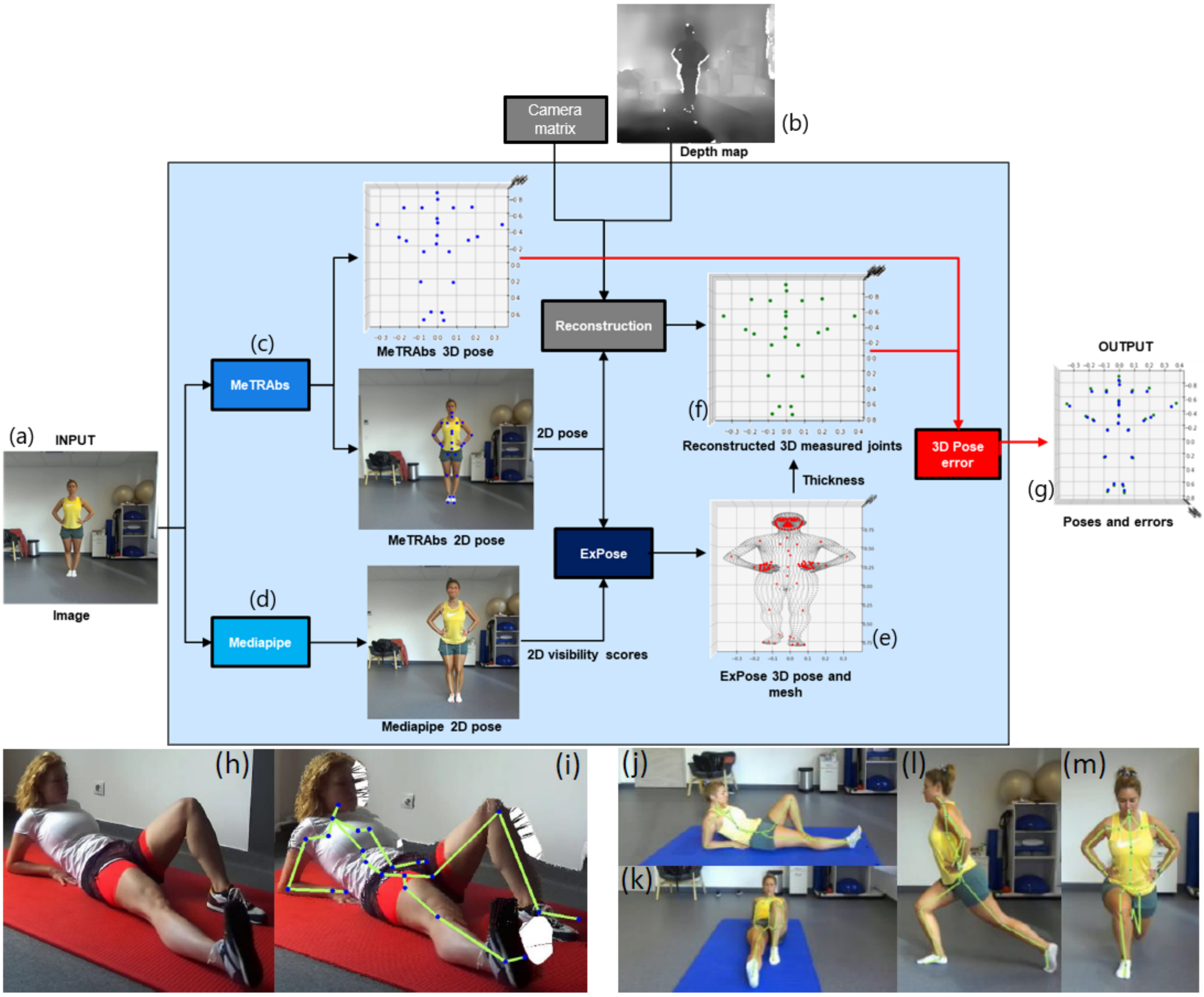


Figure 2. Body Pose module. (**a**) RGB input, (**b**) depth input, (**c**,**d**) 3D skeleton points estimated by MeTRAbs and MediaPipe, (**e**) ExPose 3D surface point estimation, (**f**) joint points estimation using Zed2 data and ExPose surface points, (**g**) joint error estimation, (**h**) large perspective distortions, (**i**) corrected image appropriate for skeleton estimation, (**j**–**m**) frames from two movements from two angles having error estimations. Best viewed by zooming in. For more details, see Section 2.4.

We proceed as follows: Section 2 details the methods we used. Experiments are contained in Section 3 and followed by the discussion (Section 4).

2. Materials and Methods

We collected video and textual data that we detail in Sections 2.1 and 2.2. Our methods include technology components and experiences gained from these data. Data (Section 2.1) contain video data and the basic expert dialogues. Section 2.3 describes our procedures for spatial tasks, including the description of the semantic and instance maps, position optimization as input to the path planning module, while Section 2.4 discusses our body pose estimation methods. Section 2.5 is about the different ways that the patient can interact with the system using videos and verbal communication. This section details the language modules for spatial navigation and for exercise related communication. An overview of the envisioned system is depicted in Figure 1.

2.1. Video Data

The videos of about 50 knee rehabilitation exercises and 400 potential error videos (about 5 to 10 for each exercise) were collected in the physiotherapy room of the Emineo Private Hospital performing knee replacements [31]. Expert physiotherapist performed the sample exercises and the erroneous ones. The 450 videos were recorded on an iPhone 11 in order to approximate home conditions. In addition, a large subset of the videos was also recorded with the ZED-2 stereo camera (https://www.stereolabs.com/zed-2/, accessed on 22 May 2022) in order to generate ground data for estimating 3D pose errors.

2.2. Text Data

2.2.1. Basic Expert Dialogue Database

We obtained from the physiotherapist answers to potential questions, e.g., during the anamnesis, error-correcting instructions for the rehabilitation exercises for all recorded videos concerning the individual errors in the 400 error videos and their proposed solutions, a dictionary that explains in-domain terminology, and dialogues concerning the level of the pain, the progression of the pain, and decisions on whether the exercise can be continued or it is better to contact the clinic.

More data were collected by crowdsourcing that we detail below.

2.2.2. Crowdsourcing User Requests

In order to extend our rule-based dialogue system (Section 2.5.5) and develop the data-driven system (Section 2.5.6), we employed the Prolific crowdsourcing platform (https://www.prolific.co, accessed on 22 May 2022) and collected dialogue data from crowd workers in addition to the data discussed in Section 2.2.1.

In order to make the data collection cost-efficient and still allow enough freedom in the dialogues, we choose a combination of the self-dialogue [32] and schema-guided [33] data collection. Our scenario involves dialogues that are task-oriented in the sense that exercises include instruction or steps with a fixed order (i.e., schema-guided). However, it also includes features of open-domain dialogues because patients can potentially comment or ask a wide variety of questions that might or might not require expert knowledge or specific treatment. We primarily aimed at collecting patient's and therapist's utterances that do not need any expert knowledge, but are tightly related to the particular exercises (we used the self-dialogue technique here).

For the purpose of our experiments, we reduced the set of distinct exercises to 12. In preliminary trials, we found that it is very difficult to explain the exercises to participants properly for the schema-guided collection. We thus fall back to a setting where a rule-generated dialog template is provided to workers and they are asked to enhance it with self-dialogue. In addition to the template, workers are provided with a video showing a correct execution, list of high-level exercise steps, and suggestions for enhancements.

Rule-Generated Templates

The templates always include two repetitions of the particular exercise. The first includes therapist's instructions for each step of the exercise, whereas the second is more loose and the therapist only sometimes encourages the patient. The patient's and therapist's turns are strictly alternating. Initially, all user turns are natural-language descriptions of events detected by the pose estimation system, e.g., "her left leg is flat on the floor". To target possible mistakes in exercise execution, we incorporate the most common problems for each exercise and sample them at random during the creation of templates. The therapist's utterances include multiple, randomly sampled synonymous variants to make the dialogues more diverse. The fully generated templates usually include around 30 turns.

Worker Enhancements

We asked workers to fill in at least 20 new turns into the given template while keeping the patient's and therapist's responses alternating. We suggested possible patient actions such as making sure to do the exercise correctly, asking questions about the exercise or body parts involved, or other comments. We explicitly prohibited the workers from introducing new exercise steps or repetitions, and from adding new turns at the beginning and end of the dialogues.

In total, we collected 243 dialogues (see Table 1 for an example) with more than 7000 distinct utterances. We reserved 24 dialogues for validation and another 24 for test purposes (i.e., exactly two dialogues for each distinct exercise).

Table 1. Part of a dialogue collected in crowdsourcing. Enhancements added by workers are marked with ✓.

Actor	Enh.	Utterance
therapist		Please set up in the starting position. Lie on your back.
patient	✓	Do I need a towel or mat to lay on for this exercise?
therapist	✓	Yes a towel or mat is a good idea to use for this exercise.
video		*Lying on her back*
therapist		Rest your arms straight on the floor by the side of your trunk and with your palms facing down.
patient	✓	Should my arms be touching the sides of my body?
therapist	✓	Yes if that helps keep them straight you may have them slightly touch side of your body.

2.3. Methods in 3D Space

Our modules use the RGBD data to, among other things, (a) build a semantic and an instance map, (b) find the optimal position and orientation of the human and the camera for a given exercise, which is used to plan the path to the optimal position, followed by the interaction needed to get there, and (c) estimate the 3D body configuration for error detection during execution, which may also involve interaction.

2.3.1. Semantic and Instance Map Module

Maps of this type may take the form of a 3D mesh of the surfaces of the objects in the environment with semantically labeled mesh points. There are other forms, e.g., a voxel map of the environment with semantically labeled voxel points, see, e.g., [13,34]. Since 3D body position and pose can be estimated (see Section 2.4) and the allocentric map can be translated and rotated, the robot can talk about the environment in terms of the egocentric coordinates of the partner. This is common in human interaction and it simplifies the interaction.

Robots have the capability of mapping unknown environments using mounted sensors, which could be used to construct 3D dense semantic models. The semantic information has a variety of uses including but not restricted to navigation, where the map is used to shift the egocentric coordinate system to other agents operating in the environment, including the human partner. These sensors include independent 2D RGB vision sensors, depth sensors, or a combination of both, such as in the case of RGBD cameras. Pre-trained semantic segmentation models such as MaskRCNN [35] are utilized to acquire the semantic labels from the RGB image sensors. Afterward, these labels can be back-projected to the reconstructed 3D point cloud at a pixel level using specialized tools such as Kimera [34] in real-time.

Improvements to Kimera's visual odometry can be established using the state-of-the-art tool UcoSlam [36]. The semantic back-projection process labels independent voxels within the environment with the appropriate categories retrieved from the 2D RGB frames and depth information. We generate the 3D semantic and instance maps from different viewpoints in real-time [13], e.g., from the viewpoint of the human partner that can simplify machine instructions, e.g., the machine may talk about *to your left/right* and so on.

To shorten the development and testing of the navigation approach, we used the Replica dataset [37] in the Habitat simulated environment [28]. Replica offers several reconstructed indoor spaces and semantically labeled objects which form the semantic and instance maps of the environment, and Habitat allows movements in these scenes.

2.3.2. Position Optimization Module

The precision of body motion error estimation depends on the distance between the relative pose of the camera and the body. The choice of the best position and pose is thus desired. Errors come from the present methods that use 2D camera recordings and estimate body landmarks in the 2D image as the first step followed by 3D shape and pose estimations. Even the highest accuracy 2D human pose estimation models can sometimes yield results with an error margin of more than 20 cm in spite of the large datasets that are now available for training deep neural networks for 3D pose estimation.The variety of parameters, e.g., shape, size, clothes, and human pose variations can corrupt the precision of pose estimation. Further errors can stem from perspective distortions, self-occlusions and occlusions by other objects. We minimize the consequences of these challenges by maximizing the visibility of body joints, that is, by optimizing the placement and relative direction of the subject and the camera. The optimization pipeline functions as follows:

Step 1: Extract landmarks from MediaPipe's human body pose estimation;

Step 2: Construct a 3D semantic representation of the environment or use one of the provided simulated environments in the Replica dataset;

Step 3: Use a shape model such as ExPose to compute the maximum width, depth, and height of the user's physical activity boundaries throughout the planned motion pattern;

Step 4: Employ the information obtained in the previous steps to optimize the position. This includes the following: (i) Generate a 2D top-down map of the environment according to maximum height constraints computed in Step 3. (ii) Search the map for positions where the planned activity can be performed by shifting and rotating the activity boundaries computed in Step 3 to fit the the environment. (iii) The optimal position is selected by maximizing the free area surrounding the activity region. (iv) Select the camera's position, where it should correspond to the required distance and angle relative to the human partner. (v) Check the camera's field of view for obstacles that may block visibility. (vi) In the case that no suitable camera position was found, select the next optimal activity placement, by decreasing the distance between the camera and the human until perspective distortion can be compensated (see later) with prescribed precision. (vii) In case no solution was found, the depth estimation might not then be precise, which should be noted. Additional constraints of minimizing occlusions—that depends on the exercise—can be added for sorting the candidate positions. We mention that the body pose estimation methods can compensate for limited occlusions [22].

2.3.3. Path Planning Module

Path planning from the actual position to the goal position is a standard technique [38] in our 2D context. A slight complication may arise from size and body shape constraints. Shape estimation is detailed below.

2.4. 3D Body Pose and Shape Estimation Module

In the case of rehabilitation and upon reaching the optimal position, the task is to perform the exercise that the system should monitor and evaluate. 3D pose estimation uses 2D camera recording and starts by 2D landmark estimation being quite precise. 3D errors are however considerably larger and depend on the relative pose between the camera and the body. We wanted to estimate the 3D error and used a 3D camera and a combination of methods to estimate the 3D error.

The body pose module enables us to monitor and estimate the error of body pose by means one of the following methods:

(a) MediaPipe produces 33 body landmarks and provides visibility scores relevant for detecting occlusions. It works in real-time even with moderate computational power, such as mobile devices.
(b) MeTRAbs estimates up to 122 joints containing spine joints that give rise to a robust pose estimation. It provides several backbone models with different speed and accuracy trade-offs.
(c) ExPose is a 3D body pose and shape estimator that optimizes SMPL-X parameters based on estimated 2D landmarks.

The body pose error estimation module (Figure 2) works as follows.

1. The Zed2 stereo camera provides the RGB pixel values (Figure 2a) and depth estimations (Figure 2b) for each frame;
2. MeTRAbs and MediaPipe estimate 2D and 3D poses. MediaPipe yields additional joint visibility scores (Figure 2c,d);
3. 2D body pose estimations from MeTRAbs and MediaPipe visibility scores of joints form the input to ExPose. Expose generates the 3D body pose and shape parameters. We refer to it as an "Avatar". (Figure 2e);
4. One can draw a line that connects any 3D skeleton landmark to the camera. One can find the Avatar's mesh point along this line that is closest to the camera. The distance between the skeleton landmark and this Avatar mesh point approximates the "thickness" of the body part as seen from the direction of camera;
5. We reconstruct a 3D pose using the measured depth data, the intrinsic camera parameters and the MeTRAbs 2D joints, and we reproject the 2D joints into 3D space.
6. We correct the data by adding the thicknesses to derive the depth now at the level of the joints (Figure 2f), an approximation of the ground truth at the joints.
7. We scale the MeTRAbs poses to our ground truth to minimize depth estimation error of MeTRAbs: we calculate the center of mass for both sets and the ratio between their third coordinates (depth) is our scaling factor.
8. We compute the pose error of the MeTRAbs by comparing its 3D pose estimation to the reconstructed measured 3D pose (Figure 2g).

Compensation for Perspective Distortions

Environment space may be limited and the optimal exercise position (Section 2.3.2) may be too close to the lens of the camera, giving rise to perspective distortion, and corrupting the weak-perspective camera model. Under such distortion, it can happen that neither MeTRAbs nor MediaPipe can estimate the joints, see Figure 2h.

The pixel coordinates after perspective projection and their weak-perspective simplification can be used to correct the corrupted data by means of the depth map and the camera intrinsic parameters for '+pushing' 3D voxels further from the camera in order to decrease the relative distances between the points and thus to decrease the perspective distortion [39]. We pushed the point cloud by 20% of the mean depth of the point cloud and also increased the focal length by the same percentage to achieve similar distance from objects on the projected image (Figure 2h,i).

Missing pixels on the projected image can be corrected from the original image using Delaunay triangulation [40]. Having these triangles and their correspondences in the original image, we can determine the affine transformation matrix between them and use the matrix to find the RGB values for the missing pixels in the original image. Such correction can improve the range of 3D body pose estimation.

2.5. Interactions

For each session, a single exercise is assumed. In the following, we describe the overall procedure common to all systems followed by the description of the individual levels, i.e., video-only (Section 2.5.3), rule-based dialogue (Section 2.5.5) and a fully data-driven dialogue (Section 2.5.6).

The envisioned framework works as follows: At the start of the interaction, the system makes sure the patient is ready and introduces the exercise. If the given patient is taking the planned exercise for the first time, a detailed description is given. This may be repeated later upon request. The system then provides step-by-step instructions (via video examples and potentially speech) and makes sure that the patient follows the instructions correctly (either through a basic feedback form, or through spoken feedback). The system is also responsible for handling errors external to the exercise, such as failure to correctly recognize the patient's pose or speech.

The patient follows the guidance but is also able to express pain, interrupt or cancel the exercise. Consequently, speech-enabled versions of the system allow additional inputs, such as clarification requests.

Present modules include a dialogue system for navigation to the optimal position. Different interaction methods can be used during the exercises. The approaches to be detailed only support a single exercise per session. Multiple exercises as well as discussing anamnesis with the patient, i.e., collecting information about the patient's condition and previous treatments, and dialogues concerning more accurate, personalized guidelines (e.g., lowering the difficulty of an exercise) are left for future work.

Interaction starts with spatial navigation. Exercise phases, such as instructions concerning the exercise, clarification requests, performance related notes and so on start with intent detection and are followed by dialogues. We go through these steps in this order.

2.5.1. Language Module for Spatial Navigation

The instruction generation phase is based on templates. We used the Room-to-Room (R2R) [41] dataset developed for visually grounded natural navigation in not-yet-seen houses and buildings. These templates are built by exploring the R2R vision-language instruction dataset and adapting it to flats. We determined salience upon testing navigation cases as detailed in Section 3.2.

2.5.2. Intent Detection in Speech

The user's communicative intentions are detected by an intent detector module, which classifies the utterances processed by the ASR system into intent categories. There are two higher-level categories: (i) the intent to navigate to the optimal position and (ii) intents related to the exercise itself.

As customary in modern dialogue-state architectures [42], the intent detector is a probabilistic classifier whose parameters are learned from a supervised dataset. Concretely, the intent detection module was implemented as a multinomial Naive Bayes classifier with bag of words feature vectors. Although Naive Bayes was mainly chosen to provide a simple baseline, the fact that it is a *generative* model has some important advantages, which it shares with more sophisticated, e.g., LSTM-based generative text classifiers [43]: (i) it is straightforward to make use of different intent priors depending on the dialogue state, (ii) it reaches its (somewhat higher) asymptotic error rate on fewer data points than discriminative models of similar complexity [43,44], and, consequently, it can be expected to perform better than comparable discriminative models when only a very small amount of training data are available.

Our small dataset of 200 utterances was manually created using the therapist-provided example dialogues. We analyzed the dialogue flow and found that the users' intentions motivating their utterances could be characterized by the following intent categories:

- YES, NO: positive or negative answer to a yes-no question;
- OK: acknowledging an exercise instruction;
- STOP_TRAINING, CONTINUE_TRAINING, END_TRAINING: requesting training to be stopped, resumed or ended;
- REPORT_PAIN: reporting pain during or after an exercise;
- UNABLE_TO_DO_EXERCISE;
- QUESTION_WHAT_IS: asking a clarification question about a term used in the instructions;

- QUESTION_REPEAT_COUNT: asking how many times the exercise should be repeated;
- REQUEST_REPEAT_INSTRUCTION: asking for the instructions to be repeated;
- NUMBER_1, . . . , NUMBER_10: answering a question with a numerically expressed degree on a scale of 1 to 10 (e.g., the degree of pain).

Although this set seems to be sufficient to cover the user intentions that play a role in the physiotherapist-provided dialogue examples and schemas, we expect that it will be extended by adding further intent categories on the basis of the user request dataset collected for the data-driven dialogue system (see Section 2.2.2). As a step in this direction, we manually categorized the user utterances in 10 dialogues randomly chosen from this dataset, and found, in line with the instructions provided to the crowdsourcing workers that about 15% of them expressed intents on the above list, and the majority consisted of questions about specific details of the current exercise (34%) and "am I doing it correctly now?" type approval requests (33%).

2.5.3. Video and Questionnaire Based Interactions

The simplest method for feedback and help may use video recording and evaluation together with a questionnaire about pain and including warning signs in case of increasing pain and/or establishing the connection with a human physiotherapist if needed. In this case, the camera-based system indicates if the patient is in the viewing angle, light condition is satisfactory and the exercise can be started according to the protocol. After the exercise, the video example for the correct execution of the exercise and the recorded performance of the patient can be shown side-by-side to the patient if needed. If the video processing system detects a patient error, a pre-programmed expert advice on how to improve performance is shown.

The level of pain can be asked. It is typical to ask the patient to grade it between 1 and 10 with 1 being no pain and 10 denoting very high pain level. A questionnaire may be used. The position and the type of the pain are to be asked from and characterized by the patient and should be evaluated by the physiotherapist if the level of pain is increasing. The advantage of this video-only interaction is that the probability of misunderstanding is low; virtually, the only possible point of failure is the video processing system, the rest of the architecture is very robust. The drawback is in the low flexibility of the interaction.

2.5.4. Exercise Description Procedure

The basic flow of the interaction is based on data supplied by a physiotherapist, who created several documents with detailed descriptions of the exercises, possible errors, questions related to the anamnesis, the vocabulary of body parts, and exercise movements.

The obtained exercise descriptions were converted into a structured format based on decision trees. The data are stored as YAML files, which are human-readable and machine-parsable at the same time. The nodes of the tree contain the conditions of the errors and the corresponding instructions for fixing them. The exercise procedures are a direct basis of rules used in the rule-based dialogue system, but they are also used as the basis for crowdsourcing data collection.

2.5.5. Rule-Based Dialogue (RBD) System

To increase flexibility of the interaction, we designed a RBD system based on expert-written exercise instructions and description of potential patient errors (see Section 2.2.1). It works on top of the basic step-by-step plan of the exercise, providing individual prescribed guidance. In addition to that, it processes events from the video error classification module and reacts with instructions for correction. In addition, the exercise can be interrupted at any time by user input (speech/intent or non-verbal).

Tables 2 and 3 show two example interactions with the RBD system. In the first example, the patient is asking for information. In the second one, the visual system detects an error and stops the exercise. As shown by these short excerpts, these dialogues are relatively rigid, but the RBD system offers more comfort than the video-only dialogue.

In particular, there is no need to fill in a pain questionnaire after each exercise or to look up the details of each exercise, and the dialogue system can stop the exercise immediately when an error is detected.

Table 2. Rule-based interaction. In this example, the patient is asking for information. Note that the system processes both speech and video events.

Agent	Event/Utterance
system	This is the Quads over a roll exercise. Please, follow my instructions. Please say “OK/yes/ready” if we can start, “no” if you want this message repeated, or “Question” if you want to ask something.
speech	OK.
system	First, lie on your back, and rest on your elbows.
video	*Lying on her back*
system	Put the ball under your operated knee.
video	*Ball under knee*
system	Flex your knee on the other side in order that the opposite foot can lie flat on the floor to protect your lumbar spine.
speech	Question.
system	OK.
speech	What is the lumbar spine?
system	The lumbar spine, commonly known as the waist, consists of the lowest 5 vertebrae.
speech	OK.
system	Stretching your thigh, push your knee-pit hard against the ball.
	…

Table 3. Rule-based interaction. In this example, the system detects an error. Note that the patient may indicate that they are aware of it and choose not to inform the therapist.

Agent	Event/Utterance
	…
system	Stretching your thigh, put your knee-pit hard against the ball.
video	*Knee-pit rises off the ball*
system	Stop, please.
video	*Patient is still*
system	Your performance needs correction. Should I tell you?
speech	No.
system	Can you fix it?
patient	No.
system	The exercise should be ended and your practice protocol should be modified. Should I write to your therapist?
patient	No.

There are, however, additional error sources, including the external noise that may lower the performance of the speech recognition system and intent detection, or user questions that do not fit the pre-programmed question database. In case of such errors, the system may always fall back to the video-only interaction and leave it up to the patient if they want to return to the RBD system.

2.5.6. Data-Driven Dialogue System

Sufficient amounts of in-domain data for training and evaluation are required by multiple parts of our system. Obtaining footage from real physiotherapy sessions is not feasible due to high cost and potentially sensitive information. Therefore, we work with a combination of examples collected with physiotherapists (exercise descriptions in Section 2.5.4, example videos in Section 2.1, and example dialogue in Section 2.2.1) and dialogue data obtained via crowdsourcing (Section 2.2.2).

Compared to the rule-based system, statistical alternatives based on fully data-driven dialogue systems can benefit from flexibility, especially in scenarios of unexpected user behaviour such as open-domain social conversation, leading to more engaging dialogues and better user experience. On the other hand, without a large dataset of real-world data, it might have problems with resolving very specific tasks or tasks requiring expert knowledge.

Baseline

Our baseline model is the DialoGPT [29] model fine-tuned with inputs consisting of (1) the history of the dialogue and (2) a description of the current exercise to be performed. The description contains the name of the particular exercise and the affected knee, i.e., left or right. We trained the baseline until convergence and used utterance dropout at training time to fight overfitting.

Exercise Tracking

To further adapt the model to our task and increase its ability to consistently track the exercise and detect user errors, we included a classifier head on top of the model to explicitly predict a binary flag marking invalid user steps, and a prefix to all system utterances (even those that do not explain any exercise step) marking the current exercise step and forcing the model to focus on the exercise progress, similarly to [45].

Chit-Chat Abilities

To retain the pretrained model's chit-chat capabilities and prevent catastrophic forgetting [46], we combined our collected dataset with the DailyDialog open-domain dataset [47]. During training, we replaced the last user utterance and the target system response with an example from DailyDialog at random.

We also experimented with data augmentation on the collected data using back-translations of the original dialogues, i.e., translating the texts to another language and back into English [48,49]. However, the machine translation models used do not work well in our specific medical domain and often confused body parts or produced inappropriate translations. Even though we filtered out only back-translations that contain exactly the same specific medical terms as the original utterances (e.g., glutes), we did not find this augmentation to be useful in our case.

3. Results

Our work aims at providing machine support for physical rehabilitation scenarios. All components can use dialogues. We treated goal-oriented conversations that include the next steps in a given state, i.e., navigation. We worked on conversations related to the exercises themselves.

We start with the results on body pose estimation of the Spatial Task Framework and then move on to the general problem of navigation, followed by other results related to rehabilitation.

3.1. 3D Body Pose and Shape Estimation

Efficient therapy needs precise body pose estimation for estimating errors and feedback on corrections. For testing our body pose module, we used a Zed2 stereo camera and recorded videos of an expert performing physical exercises for knee injury rehabilitation. Stereo cameras only calculate the depth for the points visible to both lenses. No depth information is available for areas detected by only one camera, as shown in the upper part of Figure 1: We use the method of Section 2.4 to compare our estimated body poses with the depth camera measurements. In order to do so, we recorded additional RGBD videos for the two selected samples shown in Figure 1. One of the exercises starts from a 'standing' position, the other starts from a 'sitting' position. We recorded both motion patterns from different distances (2.3, 3.2 and 4 m) and in different directions (frontal view, 45 degrees and profile view) making a total of 18 videos.

Error estimation proceeds as follows: MeTRAbs feeds joint estimations to ExPose. We found that this combination could not produce results for five videos. One of the cases is the standing exercise recorded from the farthest distance and profile view. Different causes are involved. One of them is the heavy self-occlusion of the leg that moves backwards. Another one is the lower resolution for larger distances. The sitting position is harder, for instance, only the frontal view had estimations for the farthest position as there is no occlusion in this case. Still, estimation was not feasible for the closest position. In this case, the perspective distortion prohibits the MeTRAbs estimation. We have shown (Figure 2h,i that perspective distortion can be compensated. Models were computed for closest and middle cases for profile view and for the closest case for the 45° view. We found that MeTRAbs is more precise for standing than for sitting, probably due to the differences in the number of training samples for the two cases.

Figure 3 depicts the errors for the visible body parts relevant for rehabilitation after TKR, that is, for the hip, the knee, the ankle, and the tip of the foot. The figure shows two views: the profile view and the 45° recordings.

Tables 4 and 5 present quantitative results for the visible body parts. Although the hip joint is far from the surface, thickness correction works well for the profile view. The largest error occurs for the foot landmark recorded at 45°. The error can be explained by the larger distance and the less precise MeTRAbs fit. The bending angle of the knee is one of the critical parameters of the exercise and errors are acceptable from such a distance and angle range.

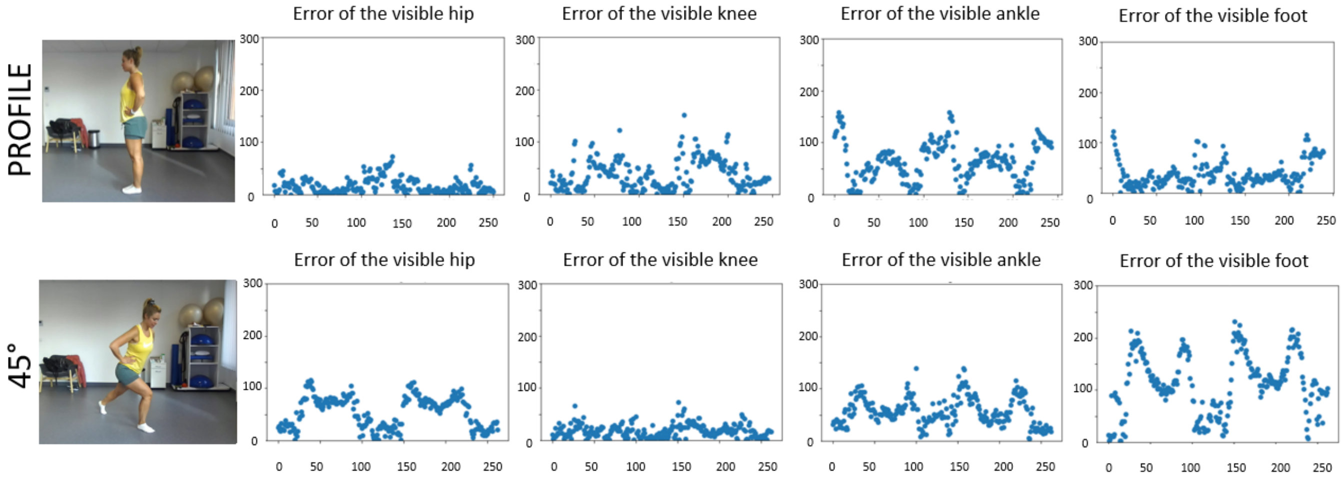


Figure 3. Errors in millimeters for the visible joints (hip, knee, ankle, foot) around the knee recorded from the profile view (upper row) and from 45° degree. The start and end body configurations are shown in the left image of the upper row. The extreme position of the exercise can be seen in the left side of the lower row. Exercise is executed twice. Error averages and standard deviations are in Tables 4 and 5. Average error is only a few centimeters, typically 5 cm or smaller, except for the foot for 45°.

Table 4. MeTRAbs 3D Body Pose error results. STD: standard deviation. Distance: 2.3 m.

Visible Joints	Pose Error	
	Profile View	**45 Degree View**
Hip	1.69 cm	5.55 cm
Knee	3.96 cm	1.93 cm
Ankle	6.67 cm	6.0 cm
Foot	4.0 cm	12.0 cm
Average (STD)	4.08 cm (2.87 cm)	6.37 cm (3.99 cm)

Table 5. MeTRAbs 3D Body Pose error results. STD: standard deviation. Distance: 3.2 m.

Visible Joints	**Pose Error**	
	Profile View	**45 Degree View**
Hip	2.28 cm	10.25 cm
Knee	2.57 cm	4.13 cm
Ankle	5.42 cm	10.1 cm
Foot	14.71 cm	9.48 cm
Average (STD)	6.25 cm (3.73 cm)	8.49 cm (3.15 cm)

Note that body parts not visible at the time can cause large errors. Such large errors are not due to MeTRAbs, as MeTRAbs estimates are conservative. Instead, they are due to the Zed2 camera used in our detection system, which is used exclusively for estimating body part thickness: Zed2 measures visible objects. Thus, only pose errors of visible joints are presented.

How can one satisfy the distance and direction constraints in an apartment? The avatar model offers the solution. The exercise can be animated if the protocol of the exercise is available. Alternatively, as we did here, the exercise can be recorded at the optimal distance and from the optimal direction by the expert. In this second case, the animation can be derived by matching the avatar to the video using MeTRAbs and ExPose. In both cases, the avatar's parameters should be modified to the parameters of the patient. The range of allowed distances and relative directions are determined by means of the avatar: error estimation should be below threshold. Then, the camera and the avatar are placed into the map as illustrated in Figure 1f. Finally, potential places in the environment are searched for using the semantic and instance maps.

The next step is to guide the patient who should put the camera to the right place and then the patient to assume the optimal position and optimal direction.

3.2. Navigation in the Spatial Task Framework

Navigation instructions depend on the current and final position. The route can be broken down into segments so that the current instruction applies to the next segment. In case of an error during segment execution, a new route is determined on the fly. Segmentation and instructions are efficient when the segments are long, i.e., a segment is about a longer path in the environment, and the instructions are short, i.e., a few words can describe the next point in the environment. These conditions may contradict to each other as the larger the distance the more complex the description of the route may become.

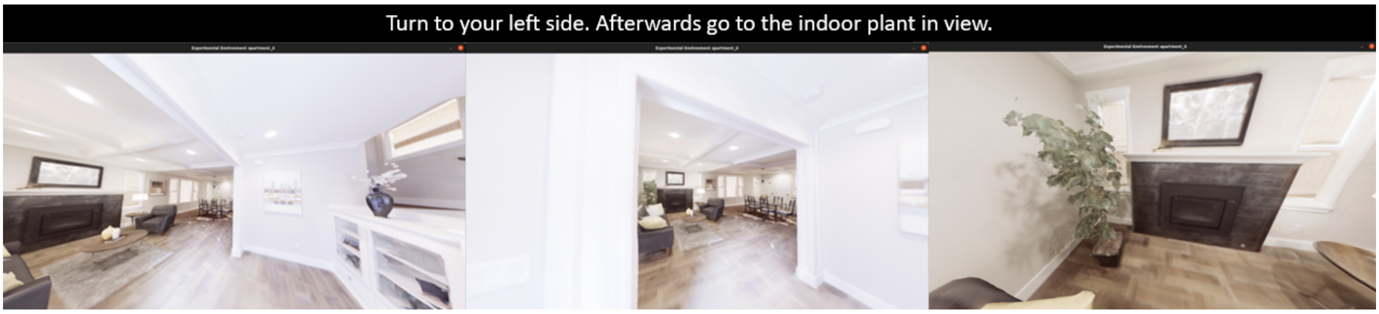
Most navigation assistance systems consider blind people or optimize for robots [50]. Our case is different: we can rely on the recognition capabilities of the patient. In order to resolve the mentioned conflict, we experimented with different methods in the Replica environment. To select robust instructions, we run several experiments in the Replica environment. Machine instructions can refer to the direction (i.e., which direction to turn) and the next position to reach.

3.2.1. Directional Instructions

We found that a combination of directional instructions using (i) clock positions and (ii) turning left, turning right, and turning around is robust: the latter can be augmented by the instructions "at 1 o'clock", "at 2 o'clock", "at 10 o'clock" and "at 11 o'clock". For example, people are less likely to make mistakes receiving the instruction "turn back and then go towards *something* at 11 o'clock", than for "turn to 5 o'clock and then go forward towards *something*", or for "turn back and then go slightly to your left towards *something*".

3.2.2. Instructions for Segment Endpoints

Note that path planning is fast and errors in execution can be quickly accounted for. We found that (i) instructions trying to follow the shortest path can be complex and easily misunderstood. (ii) Flexible path definition is preferable if it favors well-identified objects (Figure 4). (iii) The greater the angle of view of the next object, the better the choice, but (iv) the further away the object, the fewer instructions are required. (v) If there are multiple instances of a given object in the vicinity of the next position, it is preferable to choose the nearest or furthest instance. (vi) If the object is visible, then route determination is not necessary, it can be left to the human. (vii) If the object is unique, then the direction can be omitted, provided the position of object is made clear by a larger area. For example, "go to the sink in the bathroom" can work robustly if the bathroom is visible and recognizable. The following combination of instructions is effective because it designates a wide field of view within which the context helps in finding the object: 'turn around and then go to the sink in the bathroom at 11 o'clock'. (viii) Relative positions, such as "to the left of *something*" can be misleading, and it is better to use the instruction "go to *something*", and then tailor the rest of the track according to the position reached. (ix) Objects can be selected even if they are partially visible provided that the angle of view is large: instructions can rely on the user's recognition skills.



Turn to your left side. Afterwards go to the indoor plant in view.

Figure 4. (**Left**) actual position, segment starts here, instruction is received; (**Middle**) the first part of the instruction is executed; (**Right**) the second part is executed, the segment is finished.

3.2.3. Instructions for the Point of Destination

The instruction concerning the end point of the route is different because the coordinates of the end point are typically strict. In the case of exercises, the end points are the centres of larger empty spaces. We have found that the instructions 'between' and 'in the middle of something' work well. In this case, relative position can also be used. For example, '1 m in front of the television' or 'between the television and the table' are appropriate because people tend to choose the midpoint and optimise the distance. If necessary, small corrections are still possible. Egocentric information such as left and right work well when used in the form 'on the left/right', but can lead to errors when used to denote the left/right side of the object. The direction at the end point can use the clock position method and the machine can ask for the direction between 1 o'clock and 2 o'clock, for example.

3.3. Intent Detection

We evaluated our Naive Bayes intent classifier, and obtained 76.6% accuracy on the test section of our dataset (see Section 2.2.1). Performance of the intent detection module could be improved considerably by increasing the size and quality of the training corpus. Deep neural networks could also be trained on a larger corpus, which promises further improvements.

3.4. Evaluation of the Data-Driven Dialogue

We provide a basic evaluation using the BLEU score [51] to understand the outputs of our data-driven dialogue system (see Section 2.5.6) in multiple respects. BLEU is currently one of the most popular metrics for task-oriented systems, even though its correlation with human judgements has been questioned [52]. However, our dataset contains many exercise instructions produced using varying templates, and thus the BLEU score can give us a notion about the system's abilities to give proper step-by-step exercise instructions, and about its sensitivity to patient's mistakes or the video events. Human evaluation, which would guarantee accuracy, is hard to set up in the current version of the system where all components are only evaluated individually. We are planning human evaluation in future, with a full system in real-life conditions.

Table 6 summarizes the performance of the three versions of our model described in Section 2.5.6. Clearly, the baseline model cannot recognize user mistakes at all, and is also worse in case of the scores measured on the basic rule-generated instructions and the crowdsourced enhancements (questions, chit-chat, etc.). The other two versions are very close to each other in view of this comparison. However, the model trained on additional chit-chat data gained the ability to respond to a much wider variety of patient utterances in a meaningful way. See Table 7 showing a sample conversation with our best-performing model.

Table 6. Corpus evaluation. Evaluation of our data-driven models (see Section 2.5.6). Scores for each type of system responses are reported separately in columns.

Model Version		**BLEU**	
	Instructions	**Mistakes**	**Others**
Baseline	75.61	11.48	4.08
+ Exercise tracking	77.21	52.80	4.79
+ Chit-chat abilities	77.58	51.12	4.83

Table 7. Conversation with the data-driven model. A dialogue involving a single repetition of the "Quads over a roll" exercise for left leg, conducted by one of the authors.

Side	**Utterance**
system	Please, set up in the starting position. Lie on your back.
video	*Lying on her back*
system	Left leg flat on the floor.
video	*Her left leg is flat on the floor*
system	Put a roll, ball or pillow under your left knee.
patient	How big should the ball be?
system	It should be about the size of a tennis ball.
patient	Hmm, I've got only a volleyball ball. Can I use a cushion?
system	Yes, of course.
video	*Put the support under her left knee*
system	Rest on your elbows.
video	*Resting on her elbows*
system	Bend right leg 60° with your foot flat on the floor.
video	*Her right leg is bent 45° and her foot is flat on the floor*
system	Incorrect! Bend right leg 60° with your foot flat on the floor.
video	*Her right leg is bent 60° and her foot is flat on the floor*
system	Pump your left ankle.
patient	Pump? What do you mean by "pump"?
system	Move it up and down.
patient	Like a pipe?
system	Yes, exactly like a pipe.
video	*Her left ankle pumped*
system	Now the repeating part. Stretch your left thigh muscle.
patient	What is the thigh muscle?
system	The thigh muscle is the area around your thigh.
video	*Her left thigh muscle stretched*
system	Push your knee-pit hard against the support.

Table 7. *Cont.*

Side	Utterance
patient	What weather do you like most?
system	I like the cold.
patient	Cool, I like ice-cream.
system	What kind of ice cream do you like?
video	*Pushing her knee-pit hard against the support*
system	Slowly lift your left heel up in the air to straighten your leg.
patient	I like chocolate and orange.
system	What kind of chocolate do you like?
patient	The brown one.
system	What kind of orange?
video	*Her left heel lifted up in the air and her leg straighten*
system	Hold this position for 2–3 s.
video	*The position held for 2–3 s*
system	Slowly lower your left heel to the floor.
video	*Her left heel lowered to the floor in 2 s*
system	Relax your left thigh muscle.
patient	Uh, it was pretty hard to hold the position for so long.
system	I know, but you're doing great.
video	*Her left thigh muscle relaxed*
system	Great work.

4. Discussion

We considered physical rehabilitation scenarios but some components, such as spatial tasks, or body pose estimation, are more general. For example, spatial navigation may help elderly people with moderate dementia, while body pose estimation can help somebody locate smaller objects or handle larger ones.

In our studies, we focused on rehabilitation exercises following Total Knee Replacement. Out of the interaction types and their diverse complexities, this task is relatively straightforward, as it has natural segmentation, prototypical motion patterns for comparisons, assumed absence of perturbations and that it should be executed far from obstacles, so the pose can be optimized. It is still complex and calls for a number of modules, including semantic and instance maps to facilitate verbal communication, 3D body pose estimation with high precision and dialogue systems for navigation, the description of the exercise, and for explanations of the errors together with methods for correcting those.

We start by discussing our Spatial Task Framework. STF has two main components: spatial tasks in the environment, such as spatial navigation, and local body motion related tasks concerning body configuration at a given place, such as rehabilitation exercises. Manipulations also belong to STF but are out-of-scope. We continue by reviewing communication, including dialogue systems and feedback. At the end of this section, we also give an outlook for future developments.

4.1. Spatial Tasks Concerning the Environment

We have developed a navigation method that guides the partner to the optimal position. This spatial guidance system was tested and optimized taking into account the trade-off that communication should be minimal and accurate. A similar system can be developed for movement patterns, e.g., for exercises of different kinds. This latter is not treated here.

For our database, which is about knee rehabilitation, we found that the ranges of satisfactory optimal position and pose direction are restricted. However, if they are satisfied, then a small error rate can be ensured. The optimal position is to be reached and the environment may be novel to the human partner. Semantic and instance maps can solve the problem.

4.1.1. Optimal Position

Computation of the optimal position is relatively easy if the 3D mesh of the environment is available. Classical methods, including exhaustive search, can be invoked. Errors may arise for semantic reasons: the method may not know if the place having the optimal geometry is cold, or hot, or wet. Semantic and instance maps can overcome this problem and can simplify communication too.

4.1.2. Semantic and Instance Map

Semantic and instance maps, i.e., object-segmented 3D meshes with semantically labeled mesh points (note that if mesh points have proper semantic labels then the objects of the environment can be segmented from any viewpoint) are key components for many scenarios, including rehabilitation exercises, not mentioning collaborative tasks when the position and the pose of the human partner are relevant. These maps can be developed on-the-fly. Upon estimating the position and the pose of the partner, the map enables the machine to approximate what the human partner may see, i.e., to see 'through the eyes' of the partner, and to communicate according to the partner's egocentric coordinate system. This technology component needs shared object semantics as considered in [13]. In addition to egocentric frames of reference, instance maps also enable object-centric communication, and we use both: we say 'to your right', or 'to the left of the chair'. The latter can be ambiguous for more chairs, while the former needs visual information. None of this is possible without semantic and instance information.

4.1.3. Navigation to the Optimal Position

Using virtual reality, we studied navigation directions. We found that short instructions are superior to precise ones due to the following reasons: (i) long instructions may be forgotten, (ii) if the patient is visible, then it is easy to detect errors in the navigation, (iii) human players generalize well in diverse situations, and (iv) correcting instructions are easy to derive that can give rise to a precise pose.

4.2. Rehabilitation Exercises and Pose Estimation

We studied video-grounded goal-oriented dialogue systems in the context of physical rehabilitation after total knee replacement. We elaborated the evaluation of two exercises and showed in Section 3.1 that the precision of the evaluation can be sufficient if the distance and the relative direction between the body and the camera are optimized, see Tables 4 and 5. We have similar results for about half of the videos exercises. Better approximations will appear. For example, in the case of two or three cameras, high precision estimation is possible [53].

There are different solutions to overcome this obstacle. The simplest solution is to present the prototype video and the recorded video of the patient side-by-side. This way, the patient and the physiotherapist can estimate the errors for the other cases. Quantitative comparisons about the improvements are not possible in this case.

Novel methods offer improved solutions. For instance, high precision body pose estimation can be achieved. For example, two or three cameras can combine available information by means of body pose dictionary learning using deep sparse coding technology [53]. In addition, two- or more cameras can overcome problems arising from self-occlusions. Another recent approach based on neural radiance fields [54] offers improved precision.

We conclude that pose estimation technology is ripe.

4.3. Communication with the Patient

We turn to the problem of communicating with the patient. For this problem set, the video-only dialogue is available. This case assumes that the patient can detect the differences between the prototype movement and her recorded exercise. However, expert advice can overcome difficulties and can guide the attention to the relevant joints and muscles.

Both the rule-based and the data-driven dialogue systems need further developments. For example, experts can develop more sophisticated rules. Alternatively, an expansion of the rehabilitation dialogue database should improve the dialogues. We foresee a combination of the two methods due to the medical nature of the rehabilitation scenario: no supervisory error is allowed.

We consider video processing data as a valuable source of additional information for patient intent and pain detection since video recording enables non-verbal communications. Non-verbal cues go beyond patient initiated communication: they may indicate (a) intentions that will give rise to speech events later, e.g., signs of pain and (b) emerging problems that may not be recognized by the patient yet, e.g., the improper pace of the exercise, or harmful movements and potential dangers experienced by others under similar conditions. Non-verbal events can be treated as organic parts of the interaction, to which the system reacts as a human therapist would. Multimodal intent detection can make the system more pro-active and and safer for home usage; see the recent DSTC9 multimodal challenge [55] for results in this direction.

4.4. Feedback and Feedforward Help

Feedback can use the recorded video. The physiotherapist can look at the video and explain the error over the Internet. Another option is to show the original prototypical video recorded under the optimal conditions, matching the pose of the avatar to the recorded one and showing them side side-by-side.

Verbal, rule-based instructions about possible upcoming motion errors can serve the patient well. This is similar to the practice in real physiotherapy sessions, especially in group sessions: the expert mentions relevant information about the body and muscles to avoid possible upcoming errors. This method is feedforward, and it anticipates potential errors.

Repeated exercises can use the history and can combine feedback and feedforward methods. This combination can be relevant as some errors are hard to correct if the dynamics is off-track, while proactive warnings about all errors are not feasible.

4.5. Composite AI

Composite AI [1] combines deep neural networks with higher level knowledge. Deep neural networks are closer to sensors, whereas higher level knowledge deals with the outputs of the networks, combining and constraining them. Although there are shortcuts, visual processing, speech processing and speech generation assume each other in the general machine-assisted rehabilitation scenario. Position optimization, for example, exploits measured data about the errors in order to find the right position for the camera and the patient. The search is based on the semantic information derived from a 3D mesh and semantic segmentation, producing the semantic and instance maps. Instructions use the observations and the produced maps for guidance during navigation. Feedback is based on detected errors, whereas feedforward warnings can use the probabilities of potential errors, based on the quantitative error history of the patient.

Beyond the necessity and simplicity of these concepts, Composite AI has the following feature: it can take advantage of human experiences and wisdom. Rules and relations are data driven and compressed to verbal, i.e., to the symbolic level. Collected experiences may also overwrite the rules and the relations. An example is the feedforward help that comes from experience from a large population and may be modified, personalized for the sake of progress of the actual patient given the progress and the errors that occurred.

5. Conclusions

We considered rehabilitation scenarios. We showed that technology components available today can be put together in an application that combines navigation and motion patterns. The methods include visual components for the presentation of the prototypical exercises, visual processing of the ongoing episodes starting from navigation to task execution as well as object recognition, and semantic and instance map construction. The maps simplify communication as they enable the machine to see with the eyes of the human partner by estimating her position and pose in 3D, adjusting the map accordingly and communicating according to the partner's egocentric system. Thus, natural language modules appear in a prominent role.

From the point of view of the applications, there are a few, but relevant bottlenecks. One of them is the combination of dialogues using rule-based components, chit-chat components, and their combinations. Combinations could include the outcomes and experiences either as the examples to be followed, i.e., stories, or the corollaries of the joint experiences during the interaction series and to include chit-chat extensions and modulations to ease and color the situation. Evolution is fast in this domain [29,56,57] and rule-based systems can be used in the meantime.

The other bottleneck is automated speech recognition and natural language understanding in case of background noise and other speakers. This technology is evolving quickly, see, e.g., [58] and the recent publications [59,60], but further improvements are needed.

The real-time constraint and the integration of the various components are also challenging. In this respect, our example is the video-based feedback method augmented with a questionnaire that can report about pain levels and may call the expert for online monitoring and online help is a viable solution.

A wide variety of applications for home, office, and sport can be expected to appear and/or improve.

Author Contributions: Conceptualization, O.D. and A.L.; Data curation, Z.G.; Methodology, L.H., J.S., O.D. and A.L.; Software, G.B., B.C.D.S.M., A.S., D.S., J.S. and T.N.; Supervision, O.D. and A.L.; Validation, G.B., B.C.D.S.M., D.S., J.S. and T.N.; Visualization, G.B.; Writing—original draft, G.B., B.C.D.S.M., A.S., O.D. and A.L.; Writing—review and editing, G.B., B.C.D.S.M., A.S., O.D. and A.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially supported by (a) the European Commission funded project Humane AI: Toward AI Systems That Augment and Empower Humans by Understanding Us, our Society and the World Around Us (Grant No. 820437), (b) the "Application Domain Specific Highly Reliable IT Solutions" project implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the Thematic Excellence Programme TKP2020-NKA-06 (National Challenges Subprogramme) funding scheme, and (c) the Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program. This work used services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz, accessed on 22 May 2022), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101). The authors thank the above projects and Robert Bosch, Ltd. Budapest, Hungary for their generous support to the Department of Artificial Intelligence.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data was collected at the clinic and is not public.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

THR	Total hip replacement
TKR	Total knee replacement
NLP	Natural language processing
STF	Spatial task framework

YAML	Yet another markup language
RBD	Rule-based dialogue
BLEU	Bilingual evaluation understudy

References

1. Gartner Group. 5 Trends Drive the Gartner Hype Cycle for Emerging Technologies. 2020. Available online: https://www.gartner.com/smarterwithgartner/5-trends-drive-the-gartner-hype-cycle-for-emerging-technologies-2020 (accessed on 22 May 2022).
2. iHealthcareAnalyst, Inc. Global Home Rehabilitation Market $225 Billion by 2027. Available online: https://bit.ly/3Ox9WOm (accessed on 22 May 2022).
3. der Loos, V.; Machiel, H.; Reinkensmeyer, D.J.; Guglielmelli, E. Rehabilitation and health care robotics. In *Springer Handbook of Robotics*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 1685–1728.
4. Akbari, A.; Haghverd, F.; Behbahani, S. Robotic home-based rehabilitation systems design: From a literature review to a conceptual framework for community-based remote therapy during COVID-19 pandemic. *Front. Robot. AI* **2021**, *8*, 181. [CrossRef] [PubMed]
5. Yedidsion, H.; Deans, J.; Sheehan, C.; Chillara, M.; Hart, J.; Stone, P.; Mooney, R.J. Optimal use of verbal instructions for multi-robot human navigation guidance. In *International Conference on Social Robotics*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 133–143.
6. Wilson, G.; Pereyda, C.; Raghunath, N.; de la Cruz, G.; Goel, S.; Nesaei, S.; Minor, B.; Schmitter-Edgecombe, M.; Taylor, M.E.; Cook, D.J. Robot-enabled support of daily activities in smart home environments. *Cogn. Syst. Res.* **2019**, *54*, 258–272. [CrossRef] [PubMed]
7. Foley, K.T.; Luz, C.C. Retooling the health care workforce for an aging America: A current perspective. *Gerontol.* **2021**, *61*, 487–496. [CrossRef]
8. Santos, N.B.; Bavaresco, R.S.; Tavares, J.E.; Ramos, G.D.O.; Barbosa, J.L. A systematic mapping study of robotics in human care. *Robot. Auton. Syst.* **2021**, *144*, 103833. [CrossRef]
9. Spiess, A.A.F.; Skempes, D.; Bickenbach, J.; Stucki, G. Exploration of current challenges in rehabilitation from the perspective of healthcare professionals: Switzerland as a case in point. *Health Policy* **2022**, *126*, 173–182. [CrossRef] [PubMed]
10. Byron, D.; Koller, A.; Oberlander, J.; Stoia, L.; Striegnitz, K. Generating instructions in virtual environments (GIVE): A challenge and an evaluation testbed for NLG. In Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation, Arlington, VA, USA, 20–21 April 2007.
11. Anderson, P.; Chang, A.; Chaplot, D.S.; Dosovitskiy, A.; Gupta, S.; Koltun, V.; Kosecka, J.; Malik, J.; Mottaghi, R.; Savva, M.; et al. On evaluation of embodied navigation agents. *arXiv* **2018**, arXiv:1807.06757.
12. Puig, X.; Shu, T.; Li, S.; Wang, Z.; Liao, Y.H.; Tenenbaum, J.B.; Fidler, S.; Torralba, A. Watch-And-Help: A challenge for social perception and human-AI collaboration. In Proceedings of the International Conference on Learning Representations, Virtual, 3–7 May 2021.
13. Rozenberszki, D.; Sörös, G.; Szeier, S.; Lőrincz, A. 3D Semantic Label Transfer in Human-Robot Collaboration. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2602–2611.
14. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2D pose estimation using part affinity fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Venice, Italy, 22–29 October 2017; pp. 7291–7299.
15. Sun, K.; Xiao, B.; Liu, D.; Wang, J. Deep high-resolution representation learning for human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5693–5703. [CrossRef]
16. Chen, Y.; Tian, Y.; He, M. Monocular human pose estimation: A survey of deep learning-based methods. *Comput. Vis. Image Underst.* **2020**, *192*, 102897. [CrossRef]
17. Bazarevsky, V.; Grishchenko, I.; Raveendran, K.; Zhu, T.; Zhang, F.; Grundmann, M. BlazePose: On-device Real-time Body Pose tracking. *arXiv* **2020**, arXiv:2006.10204.
18. Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; Black, M.J. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.* **2015**, *34*, 1–16. [CrossRef]
19. Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A.A.; Tzionas, D.; Black, M.J. Expressive body capture: 3D hands, face, and body from a single image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 10975–10985.
20. Choutas, V.; Pavlakos, G.; Bolkart, T.; Tzionas, D.; Black, M.J. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 20–40.
21. Sárándi, I.; Linder, T.; Arras, K.O.; Leibe, B. MeTRAbs: Metric-Scale Truncation-Robust Heatmaps for Absolute 3D Human Pose Estimation. *IEEE Trans. Biom. Behav. Identity Sci.* **2020**, *3*, 16–30. [CrossRef]
22. Lugaresi, C.; Tang, J.; Nash, H.; McClanahan, C.; Uboweja, E.; Hays, M.; Zhang, F.; Chang, C.L.; Yong, M.G.; Lee, J.; et al. Mediapipe: A framework for building perception pipelines. *arXiv* **2019**, arXiv:1906.08172.
23. Mangal, N.K.; Tiwari, A.K. A Review of the Evolution of Scientific Literature on Technology-assisted Approaches using RGB-D sensors for Musculoskeletal Health Monitoring. In *Computers in Biology and Medicine*; Elsevier: Amsterdam, The Netherlands, 2021; p. 104316. [CrossRef]

24. Buhagiar, M.A.; Naylor, J.M.; Harris, I.A.; Xuan, W.; Kohler, F.; Wright, R.; Fortunato, R. Effect of inpatient rehabilitation vs a monitored home-based program on mobility in patients with total knee arthroplasty: The HIHO randomized clinical trial. *JAMA* **2017**, *317*, 1037–1046. [CrossRef]
25. Liao, Y.; Vakanski, A.; Xian, M. A Deep Learning Framework for Assessing Physical Rehabilitation Exercises. *IEEE Trans. Neural Syst. Rehabil. Eng.* **2020**, *28*, 468–477. [CrossRef] [PubMed]
26. Boyer, P.; Burns, D.; Whyne, C. Out-of-Distribution Detection of Human Activity Recognition with Smartwatch Inertial Sensors. *Sensors* **2021**, *21*, 1669. [CrossRef] [PubMed]
27. Muoio, D. Hinge Health Now Valued at $3B Following $300M Series D. Available online: https://www.mobihealthnews.com/news/hinge-health-now-valued-3b-following-300m-series-d (accessed on 22 May 2022).
28. Savva, M.; Kadian, A.; Maksymets, O.; Zhao, Y.; Wijmans, E.; Jain, B.; Straub, J.; Liu, J.; Koltun, V.; Malik, J.; et al. Habitat: A platform for embodied AI research. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 9339–9347.
29. Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; Dolan, B. DIALOGPT: Large-Scale Generative Pre-training for Conversational Response Generation. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Online, 5–10 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 270–278.
30. Müller, M.; Koltun, V. Openbot: Turning smartphones into robots. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 9305–9311.
31. Gunn, F.; Walmsley, P. 542 The Attune Total Knee Replacement: Early Clinical Performance Versus an Established Implant At 3 Years Post-Surgery. *Br. J. Surg.* **2021**, *108*, znab134.562. [CrossRef]
32. Byrne, B.; Krishnamoorthi, K.; Sankar, C.; Neelakantan, A.; Goodrich, B.; Duckworth, D.; Yavuz, S.; Dubey, A.; Kim, K.Y.; Cedilnik, A. Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 3–7 November 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 4516–4525. [CrossRef]
33. Mosig, J.E.M.; Mehri, S.; Kober, T. STAR: A Schema-Guided Dialog Dataset for Transfer Learning. *arXiv* **2020**, arXiv:2010.11853.
34. Rosinol, A.; Abate, M.; Chang, Y.; Carlone, L. Kimera: An open-source library for real-time metric-semantic localization and mapping. In Proceedings of the 2020 IEEE International Conference on Robotics and Automation (ICRA), Philadelphia, PA, USA, 23–27 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1689–1696.
35. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R.B. Mask R-CNN. *arXiv* **2017**, arXiv:1703.06870.
36. Muñoz-Salinas, R.; Medina-Carnicer, R. UcoSLAM: Simultaneous localization and mapping by fusion of keypoints and squared planar markers. *Pattern Recognit.* **2020**, *101*, 107193. [CrossRef]
37. Straub, J.; Whelan, T.; Ma, L.; Chen, Y.; Wijmans, E.; Green, S.; Engel, J.J.; Mur-Artal, R.; Ren, C.; Verma, S.; et al. The Replica Dataset: A Digital Replica of Indoor Spaces. *arXiv* **2019**, arXiv:1906.05797.
38. Gasparetto, A.; Boscariol, P.; Lanzutti, A.; Vidoni, R. Path planning and trajectory planning algorithms: A general overview. In *Motion and Operation Planning of Robotic Systems*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 3–27.
39. Hartley, R.; Zisserman, A. *Multiple View Geometry in Computer Vision*; Cambridge University Press: Cambridge, UK, 2003.
40. Ito, Y. Delaunay Triangulation. In *Encyclopedia of Applied and Computational Mathematics*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 332–334.
41. Anderson, P.; Wu, Q.; Teney, D.; Bruce, J.; Johnson, M.; Sünderhauf, N.; Reid, I.D.; Gould, S.; van den Hengel, A. Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. *arXiv* **2017**, arXiv:1711.07280.
42. McTear, M. *Conversational AI: Dialogue Systems, Conversational Agents, and Chatbots*; Morgan & Claypool Publishers: San Rafael, CA, USA, 2020.
43. Yogatama, D.; Dyer, C.; Ling, W.; Blunsom, P. Generative and Discriminative Text Classification with Recurrent Neural Networks. *arXiv* **2017**, arXiv:1703.01898.
44. Ng, A.Y.; Jordan, M.I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes. In *Advances in Neural Information Processing Systems*; MIT Press: Cambridge, MA, USA, 2002; pp. 841–848.
45. Shalyminov, I.; Sordoni, A.; Atkinson, A.; Schulz, H. Hybrid Generative-Retrieval Transformers for Dialogue Domain Adaptation. *arXiv* **2020**, arXiv:2003.01680.
46. McCloskey, M.; Cohen, N. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *Psychol. Learn. Motiv. Adv. Res. Theory* **1989**, *24*, 109–165. doi: 10.1016/S0079-7421(08)60536-8. [CrossRef]
47. Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; Niu, S. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Taipei, Taiwan, 27 November–1 December 2017; Asian Federation of Natural Language Processing: Taipei, Taiwan, 2017; pp. 986–995.
48. Sennrich, R.; Haddow, B.; Birch, A. Improving Neural Machine Translation Models with Monolingual Data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), Berlin, Germany, 7–12 August 2016; pp. 86–96.
49. Edunov, S.; Ott, M.; Auli, M.; Grangier, D. Understanding Back-Translation at Scale. In Proceedings of the 2018 EMNLP, Brussels, Belgium, 31 October–4 November 2018; pp. 489–500.

50. Mousavian, A.; Toshev, A.; Fišer, M.; Košecká, J.; Wahid, A.; Davidson, J. Visual representations for semantic target driven navigation. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA), Montreal, QC, Canada, 20–24 May 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 8846–8852.
51. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; Association for Computational Linguistics: Stroudsburg, PA, USA, 2002; pp. 311–318. [CrossRef]
52. Liu, C.W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; Pineau, J. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–4 November 2016; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 2122–2132. [CrossRef]
53. Dabhi, M.; Wang, C.; Saluja, K.; Jeni, L.A.; Fasel, I.; Lucey, S. High Fidelity 3D Reconstructions with Limited Physical Views. In Proceedings of the 2021 International Conference on 3D Vision (3DV), Virtual, 1–3 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1301–1311.
54. Zhan, Y.; Li, F.; Weng, R.; Choi, W. Ray3D: Ray-based 3D human pose estimation for monocular absolute 3D localization. *arXiv* **2022**, arXiv:2203.11471.
55. Gunasekara, C.; Kim, S.; D'Haro, L.F.; Rastogi, A.; Chen, Y.N.; Eric, M.; Hedayatnia, B.; Gopalakrishnan, K.; Liu, Y.; Huang, C.W.; et al. Overview of the Ninth Dialog System Technology Challenge: DSTC9. *arXiv* **2020**, arXiv:2011.06486.
56. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
57. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
58. Gabbay, A.; Shamir, A.; Peleg, S. Visual Speech Enhancement. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 1170–1174. [CrossRef]
59. Gao, R.; Grauman, K. VisualVoice: Audio-Visual Speech Separation with Cross-Modal Consistency. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
60. Montesinos, J.F.; Kadandale, V.S.; Haro, G. VoViT: Low Latency Graph-based Audio-Visual Voice Separation Transformer. *arXiv* **2022**, arXiv:2203.04099.