Intelligent Service Robotics (2025) 18:613–629
https://doi.org/10.1007/s11370-025-00603-3

ORIGINAL RESEARCH PAPER

Dr.VCoach: employment of advanced deep learning and human–robot interaction for virtual coaching

Nino Cauli[1] · Diego Reforgiato Recupero[1]

Received: 22 April 2024 / Accepted: 6 April 2025

Abstract
Ensuring healthy aging and physical fitness is crucial for leading a happy, independent, and fulfilling life. Daily gentle gymnastic training at home can significantly contribute to maintaining physical health among older adults. Virtual and robotic coaches serve as valuable tools to assist elders during their daily exercise routines. In this paper, we introduce the design of a comprehensive robotic coach capable of guiding patients throughout their training sessions. Additionally, we propose an innovative synthetic dataset generator to train the video action recognition module of the robot effectively. To evaluate the efficacy of our robotic coach, we conducted tests with nine patients aged over 50, who were asked to complete a system usability scale (SUS) survey.

Keywords Robotic coach · Active aging · Video action recognition · Transformers · Human robot interaction · Assistive robotics

1 Introduction

The world we live in is becoming increasingly frenetic with rapid population growth and the proliferation of densely populated cities worldwide. Job opportunities are abundant in these urban centers, leading many people to settle in metropolitan areas. However, city life often fosters a sedentary lifestyle, with individuals spending much of their day seated at desks or commuting between home and work. Conversely, advancements in healthcare, reduced family sizes, and improved access to food have contributed to longer life expectancies and a growing percentage of elderly individuals in the global population.

Maintaining independence throughout life is crucial for older individuals. However, aging often leads to a gradual decline in physical and mental capabilities, necessitating external support for daily activities. Engaging in a healthy and active lifestyle, including regular physical activity and walking, can delay the need for assistance later in life. Unfortunately, the sedentary nature of modern life necessitates frequent physical training sessions for older adults to maintain optimal activity levels. To ensure safety, motivation, and effective training, older adults often require the guidance of a personal therapist or trainer to schedule, monitor, and supervise their exercise routines.

The high number of elders in need of assistance in their daily training routine has led to an increased demand for dedicated personal therapists. Furthermore, some elders are unable to perform their daily training at care center facilities, requiring the therapist to visit their homes for sessions. These factors limit the number of patients that a single therapist can accommodate, resulting in a shortage of available personal therapists or trainers.

In the last decade, amid the Fourth Industrial Revolution (Industry 4.0), there has been rapid technological growth and transformation driven by increased interconnectivity and the advent of smart automation. Industries and healthcare have been particularly impacted, now relying on advanced technologies such as big data, cloud computing, deep learning, the Internet of Things (IoT), mobile devices, and robotics. Smart healthcare (s-health) has emerged as a paradigm, leveraging intelligent models, smart sensors, cloud services, and communication networks to offer context-aware healthcare services. These technological advancements provide significant support to medical personnel by automating monitoring

✉ Diego Reforgiato Recupero
diego.reforgiato@unica.it

Nino Cauli
nino.cauli@unica.it

[1] Department of Mathematics and Computer Science, University of Cagliari, Via Ospedale 72, 09124 Cagliari, Sardinia, Italy

tasks and enabling the storage of patients' medical histories in the cloud. With s-health and robotics systems, therapists can remotely track patients' training progress, allowing them to manage a larger number of patients concurrently.

Researchers are striving to automate teaching and monitoring processes by proposing virtual coaches capable of suggesting exercises, monitoring execution, and sending results to doctors [1–3]. While a computer, screen, and camera suffice to implement a virtual coach, humanoid robots offer a better solution. Interacting with a dynamic and expressive robot is more engaging than merely viewing a computer screen. A humanoid robot can demonstrate exercises, making movements easier for the elderly to understand.

In this paper, we introduce a robotic coach, Dr.VCoach, designed to guide elders in their daily training sessions to maintain physical fitness and age actively. Dr.VCoach can verbally interact with users, manage training schedules for multiple users, demonstrate exercises, monitor exercise execution, and evaluate user performance. We present the design and implementation of the complete robotic coach system and evaluate it on nine elder subjects in their homes. Each subject completed a system usability scale (SUS) survey. Additionally, we created a video action recognition dataset tailored for training models to recognize gentle gymnastic exercises and augmented it using a novel synthetic video generator.

Therefore, the contributions of this paper are as follows:

1. Definition of a new set of exercises to promote active aging.
2. Creation of a novel dataset for action prediction augmented in simulation using domain randomization, along with the release of a synthetic video creation tool.
3. Design and deployment of an affordable robotic coach tested on patients in a real-life scenario.

In this paper, we focus on the implementation of a robotic coach and its interaction with elderly patients. The integration of a smart healthcare cloud database to store patient progress for remote evaluation by therapists was not implemented. We have decided to leave this aspect for future work.

This work is part of a European project called DrVCoach, which aims to develop a robotic coach for monitoring and motivating seniors to perform a daily gentle physical exercise routine to maintain physical fitness as they age.[1]

The remainder of this paper is organized as follows: Section 2 reviews the state-of-the-art on video data augmentation, and virtual and robotic coach architectures. Section 3 provides a detailed explanation of our robotic coach design and the data augmentation system. Sections 4 and 5 present the implementation details of the synthetic video generator and the robotic coach system, respectively. Section 6 analyzes test results on the data generator and the robotic coach. Finally, Section 7 draws conclusions and future directions where we are headed.

2 Related work

In this section, we present the progress achieved by researchers thus far in developing systems designed to assist patients in their daily physical training routines and rehabilitation sessions. The objective of these systems is to automate either part or the entirety of the training or rehabilitation session, ranging from machine–human interaction to the monitoring and evaluation of exercises. Our analysis of the state-of-the-art does not solely focus on complete virtual coaching systems; instead, we also delve into the subproblem of video monitoring systems. We have divided our study into two parts: the first subsection focuses on virtual coaching systems and assistive robotics, while the second subsection delves deeper into existing video action recognition models and related data augmentation techniques.

2.1 Virtual coaching and assistive robotics

Physiotherapy, rehabilitation, and exercises promoting active aging must be regularly performed by elderly patients. Due to their advanced age, they may not always be able to undergo daily training sessions at care center facilities. In such cases, home-based training becomes the preferred solution. However, since therapists cannot always be present at the patients' homes, automated social and assistive technology becomes invaluable. The automation of monitoring, in particular, has been extensively studied by researchers, and several solutions are now available [4, 5]. Various types of sensors are available for automated monitoring in the healthcare domain, including biosignal sensors integrated within wearable devices, air quality sensors, acoustic sensors, cameras (RGB, infrared, and depth), among others [6].

Cameras, in particular, are powerful tools capable of capturing body features and motion. Supported by machine learning models, RGB cameras can precisely track the location of bodies and objects, monitor their skeletal pose, and detect their motion. External cameras offer the advantage of being noninvasive compared to wearable sensors. For example, Lee et al. [4] presented a model for exercise recognition and evaluation for post-stroke rehabilitation, utilizing threshold models with binary classifications. The model input was data captured by a Kinect sensor (Depth-RGB camera). In a subsequent work [5], the same authors presented an interactive approach that combines machine learning and rule-based models to automatically assess a patient's rehabilitation exercise and customize personalized corrective feedback.

[1] https://drvcoach.unica.it/.

Encouraging and stimulating the patient during the training session is essential. Virtual and augmented reality can be appropriate technologies for realizing virtual coaching systems. Mostajeran et al. [7] investigated the acceptance of an AR coaching system for balance training, which can be performed at home. The results suggest that older adults find the system encouraging and stimulating. However, the usability of the AR system showed a significant negative correlation with participants' age. Elderly patients may find it more difficult to empathize with a virtual avatar compared to a physical one, such as a robot. Several researchers have conducted studies to analyze the level of acceptance of socially assistive robots by elderly patients in rehabilitation and active aging training scenarios [8–11]. These studies examined the interaction between socially assistive robots and older patients [8, 10], integration of ICT and robotic technologies to assist nurse practitioners [9], and elderly people's social perception of human versus robotic coaches in the context of an active and healthy aging program [11]. All authors reported a positive reception by patients for assistive robots.

Rehabilitation and assistive robotics encompass a range of topics, including exercise training robots, smart prostheses and orthoses, monitoring for rehabilitation, and robotic smart home technologies. In their work [12], the authors present a comprehensive survey on rehabilitation robotics for persons seeking to recover physical, social, communication, or cognitive function, and assist persons who have a chronic disability in accomplishing activities of daily living.

Initial studies on assistive robotics for rehabilitation and physical training used mobile robots to monitor, assist, encourage, and socially interact with patients [13–15]. While a mobile robot's physical embodiment positively influences human task-related behavior, an anthropomorphic appearance would enhance interaction with human subjects, conveying a higher level of empathy.

Initially, humanoid socially assistive robots were adopted to assist patients in physical exercises in seated scenarios with limited motion to their upper body. Fasola et al. [16] proposed a socially assistive robot designed to engage elderly users in physical exercise. During this study, older adults showed a preference for the physically embodied robot coach over a virtual coach. Similarly, Swift et al. [17] studied an autonomous socially assistive robot (SAR) coach that investigates the effect on individuals post-stroke in a seated reaching task.

With advancements in robotics, recent studies tend to use full-body humanoid robots for assistance in rehabilitation and physical exercises. In their work [18], Nguyen et al. designed and implemented a humanoid robot coach capable of demonstrating rehabilitation exercises to patients, observing a patient's exercises, providing feedback to improve performance, and offering encouragement. The authors used imitation learning techniques based on Gaussian mixture models, making the robot easily programmable by medical experts without specific robotics knowledge. More recently, Lee et al. [19] used the Nao robot to coach, monitor, and correct post-stroke patients during their rehabilitation exercises. They studied the acceptance of the robotic coach by both patients and therapists, before and after testing the system. The model used for monitoring and evaluating the exercises performed by the patients was based on skeletal extraction and pose detection from video extracted from a Kinect sensor.

Our study is aligned with the work of Lee et al. [19], but we focus our attention on implementing a robotic coach to promote active aging instead of focusing solely on post-stroke rehabilitation. To our knowledge, our work is the first full implementation of a robotic coach using a state-of-the-art model based on vision transformers (ViT) to monitor physical exercises performed by elderly patients in realistic scenarios. Compared to similar recent studies adopting D-RGB cameras [16, 18, 19], our system uses affordable RGB webcams to monitor and analyze the patients' training sessions. This solution was made possible thanks to the generation of a synthetic augmented dataset for training.

Table 1 summarizes a comparison of the main features of our approach against state-of-the-art systems.

2.2 Video action recognition and data augmentation

Over the past decade, CNNs have become the de facto standard for image recognition [20–22]. Nevertheless, in the last 3 years, models based on ViT architecture [23] are quickly replacing CNNs [24–27]. For a comprehensive review of video transformers, please refer to the work of Selva et al. [28].

While models designed for image analysis can be used for video analysis, it is essential to consider the temporal dimension inherent in videos. For this reason, state-of-the-art models for video analysis are designed to capture temporal information. These models include optical flow-based methods [21], 3D CNNs [20], recurrent neural networks (RNN) [22, 29], and ViT with space-time attention [25].

CNNs and ViT models require a substantial amount of data for successful training. Several data augmentation techniques for image analysis have been introduced and well-presented in recent review papers by Shorten et al. and Khalifa et al. [30, 31]. These surveys provide a comprehensive overview of image data augmentation, covering basic image manipulations such as geometrical and color space transformations, kernel filters, noise injection, mixing images, and random erasing. They also delve into more recent DL approaches, including feature space augmentation, adversarial training, GAN-based techniques, and neural style transfer.

A different approach involves the generation of synthetic images using simulators capable of emulating the appear-

Table 1 Comparison of our solution with other state-of-the-art systems

Virtual and robotic coach systems						
System	Case study	Coach platform	Sensors	Action Monitoring Model	Full body motion analysis	Code available
Gockley and Matari [14]	Stroke rehabilitation	Mobile robot (Pioneer 2-DX)	Inertial measurement units	No real-time monitoring	No (upper-limb)	No
Mataric et al. [15]	Stroke rehabilitation	Mobile robot (Pioneer 2-DX)	Inertial measurement units	Threshold on inertial sensor raw data	No (upper-limb)	No
Fasola et al. [16]	Elders physical training	Bandit (Humanoid upperbody on a Pioneer 2-DX)	RGB camera	Image segmentation and rule-based models	No (upper-limb)	No
Swift et al. [17]	Stroke rehabilitation	Bandit (Humanoid upperbody on a Pioneer 2-DX)	RGB, Kinect, and IMU	No real-time monitoring	No (upper-limb)	No
Nguyen et al. [18]	Rehabilitation healthcare	Humanoid robot (Poppy)	Kinect (RGB-D)	Imitation learning on Gaussian Mixture Models	Yes	No
Lee et al. [4]	Stroke rehabilitation	Computer screen	Kinect (RGB-D)	Threshold models with binary classifications	Yes	No
Lee et al. [5]	Stroke rehabilitation	Humanoid robot (Nao) and tablet	Kinect (RGB-D)	Machine learning and rule-based models	No (upper-limb)	No
Mostajeran et al. [7]	Balance Training in an Aging Population	Meta 2 head-mounted display (HMD)	Kinect (RGB-D)	No real-time monitoring	Yes	No
Lee et al. [19]	Stroke rehabilitation	Humanoid robot (Nao) and tablet	Kinect (RGB-D)	Skeletal extraction and pose detection	Yes	No
Dr.Vcoach (our solution)	Elders physical training	Humanoid robot (Nao) and laptop	RGB camera	Timesformer (Vision Transformers)	Yes	Yes

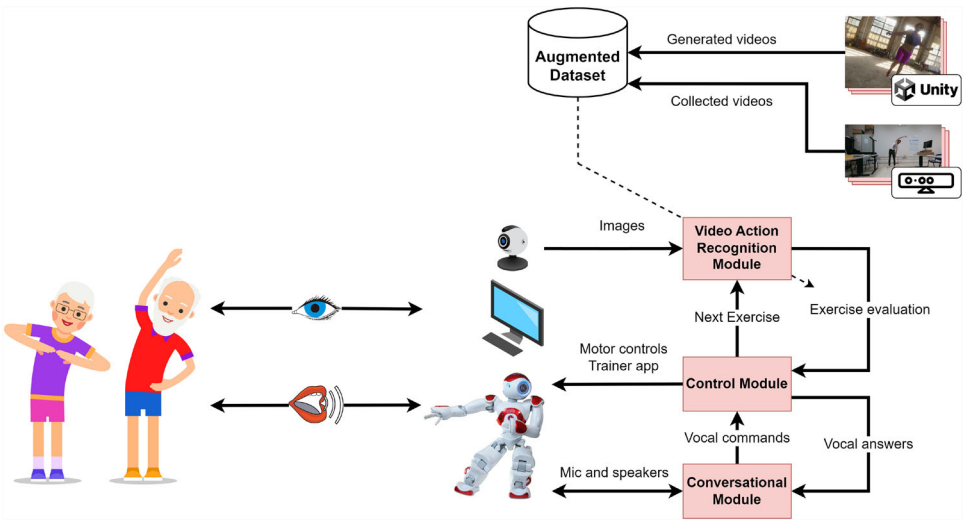

Fig. 1 System architecture

ance and physics of the real world. Game engines are often employed for generating new images due to their powerful graphic and physics engines [32]. Game engines permit the creation of synthetic images with high variability in the depicted scene, such as object positions, background, illumination, and camera positions. The concept of producing a synthetic dataset with high variability in image appearance is the core idea behind the domain randomization (DR) paradigm introduced by Tobin et al. [33]. In DR, the simulated scene parameters need to be highly randomized to generate images that fully cover the data distribution to be learned. When dealing with videos, it is essential to consider the correlation between time and space, not only in the design of learning models, but also in the generation of training datasets. In simulations, the physical interaction between objects (e.g., rigid bodies interaction and gravity), their motions, and the animation of subjects in the scene become crucial aspects for the generation of synthetic videos [34].

While several large RGB video datasets for generic action recognition tasks already exist [35–43], for more specific action recognition tasks, additional data may need to be collected. One solution is fine-tuning the DL model on these large generic video datasets. A different approach to address the shortage of data is the use of data augmentation techniques specific to videos. The recent survey titled "Survey on videos data augmentation for deep learning models" [34] provides a deep analysis of video data augmentation through simulation.

In contrast with the mentioned approaches, in this paper, we introduce a new synthetic video generator specifically designed for action recognition. While some existing solutions address object detection in static images, such as the Unreal Engine 4 plugin "NVIDIA Deep learning Dataset Synthesizer (NDDS)" [44], and others offer tools for generating synthetic videos for action recognition, such as ElderSim, a platform for synthetic data generation focusing on human action recognition within house interiors [45], none of these options provides the level of flexibility and randomization required for our augmented action recognition dataset.

3 Background

The work described in this paper focuses on the implementation of a robotic coach, named "Dr.VCoach," designed to assist elderly individuals during their daily physical training, as depicted in Fig. 1. The capabilities of the robot include:

1. Understanding verbal commands from the user through a conversational module.
2. Defining the sequence of exercises and presenting them to the user with verbal descriptions and self-performance via a control module.
3. Monitoring the user's performance during exercises using RGB cameras, identifying errors, and offering corrections through a video action recognition module.

While various datasets for video action recognition are available, none fully encompass the movements and labels required for this work. Consequently, we generated a new dataset by capturing RGB videos of individuals performing gentle gymnastic exercises in a laboratory setting. This core dataset underwent augmentation in simulation using a synthetic image generator implemented in Unity. Simulated avatars, animated with the original recorded data, were subjected to randomization in motion, background, lighting, and camera position, resulting in a significantly expanded dataset. This versatile tool serves a dual purpose: it can augment preexisting datasets, enriching them with diverse examples, or create entirely new datasets tailored to specific research needs.

To train and test our system, we selected, in consultation with a professional personal trainer, 11 gentle gymnastic exercises. Gentle gymnastics represents a form of physical activity that involves slow and progressive movements designed to mobilize the entire body. These exercises are suitable for individuals with varying levels of training, including sedentary office workers, older adults, and athletes. Regular participation in gentle gymnastics offers several advantages, including weight management, posture enhancement, and muscle toning. For older adults, a daily routine of gentle gymnastic exercises can contribute to maintaining better physical fitness. See Sect. 4 for a description of the exercises.

Initially, we collected a dataset of 1647 videos of 15 subjects performing the 11 exercises in front of a camera. Subsequently, we utilized our synthetic video generator to produce a dataset of 5000 videos depicting a human avatar executing one of the 11 exercises based on the collected video animations. The final novel augmented dataset comprises 6647 videos, obtained by combining the 5000 videos generated with the synthetic video generator with the 1647 videos collected in the laboratory.

The augmented dataset was utilized to train a state-of-the-art action recognition model (Timesformer [25]). This model underwent testing on a dataset of 239 videos featuring three subjects performing the 11 exercises outdoors. Additionally, the trained model was integrated into the video action recognition module of the final robotic coach system.

The system was implemented on Zora [46], a Nao [47] robot with a software layer designed to facilitate its use by individuals without expertise in information and communication technology (ICT). The Nao robot is commonly employed in studies involving assistive robotics, with a focus

on social interactions and affectivity. In medical and physical intervention contexts, the Nao robot is primarily utilized as a motivator or demonstrator. The Nao robot offers several advantages as a robotic coach system. First, its humanoid body makes it highly suitable for demonstrating physical exercises to patients. Second, its human-like appearance and appealing design enhance social acceptability, improving interactions with patients. Additionally, the Nao robot is more affordable than larger and more complex humanoid robots on the market, making it a practical choice for healthcare facilities and private users. Furthermore, Nao robots are widely used in academic and medical settings, fostering a large and active user community for support and collaboration.

The Aldebaran Choregraphe Suite [48] was utilized to program the 11 exercises on the robot. For the conversational module, we employed a question-answering model based on ChatGPT function calling. To monitor the actions of the subjects, we utilized the Timesformer action recognition model trained on the augmented dataset. The final robotic coach system was capable of initiating a training session with a human subject, proposing and demonstrating the training routine for the session, and monitoring the correct execution of the exercises by the subject (recognizing the action performed and indicating when the wrong exercise was being executed). We conducted tests with nine subjects and administered a survey to gather feedback on their experience interacting with the robotic coach.

4 Dataset generation

To address the common scarcity of data for video action recognition, we developed a video generator capable of producing synthetic videos featuring a subject performing specific actions. Our generator is implemented using the Unity [49] game engine along with the Perception [50] and Synthetic Humans [51] packages. Unity is distinguished by its user-friendly interface, comprehensive and practical C# scripting API, cross-platform building options, a large online asset store, and a free-to-use plan for projects without revenue. Moreover, Unity offers several plugins specifically developed for researchers in computer vision, artificial intelligence, and robotics. The Unity Perception Package is a toolkit for generating synthetic datasets for computer vision, offering a set of predefined and customizable scene randomizers, automatic labeling tools, and a C# library to customize all the parameters of the data generator. Synthetic humans, on the other hand, is a plugin used to procedurally generate and place human avatars in a virtual scene.

The videos generated by our tool feature a virtual humanoid avatar placed in a highly randomized scene, with the avatar performing one action randomly selected from a pool of preselected actions. The pool of actions used in our tests was derived from real videos using the pose landmark detection pretrained model from MediaPipe Solutions [52], a suite of libraries, models, and tools for applying AI and ML to vision, text, and audio analysis. We created scripts to extract the 3D body pose from videos of a human subject performing gentle gymnastic exercises, which we later used to generate Unity animation clips for our generator.

Specifically, our MediaPipe script extracts the skeletal pose of the subject for each video frame, representing it as joint positions and orientations in the camera's reference frame using a variant of BlazePose [53]. We then developed a C# script that first transforms the joint positions and orientations from the camera's reference frame to Unity's scene reference frame. The transformed pose is subsequently used to generate avatar animations within Unity.

Figure 2 illustrates several frames rendered by our synthetic video generator. These scenes depict a human avatar performing an exercise, set against various indoor or outdoor environments. The background is created using a skybox with a randomly selected HDRI texture, playing a crucial role in the scene's global illumination. This is complemented by the inclusion of a directional light with adjustable color and intensity. The virtual camera consistently faces the avatar, but we introduce randomization in various aspects, including its distance from the avatar, vertical and horizontal translations, longitudinal axis rotation, and 3D spatial positioning. Additionally, the avatar itself is procedurally generated using the Synthetic Humans Unity package, which can create human avatars with randomized features such as age, gender, ethnicity, height, weight, and clothing, derived from available asset pools.

For each generated video, the synthetic video generator provides the following outputs:

1. A .png image for each frame of the video. The resolution of the images is an adjustable parameter of the generator.
2. A .csv file for each frame that includes information on various labels automatically calculated by the generator. This information encompasses the name of the action, 3D camera pose, 3D global pose of each joint of the avatar, 2D pose of each joint in camera space, and avatar metadata (age, ethnicity, etc.).

The synthetic video generator and the MediaPipe scripts for extracting skeletal animations from videos and converting them into Unity animation clips are accessible online through the provided link in the footnote.[2]

[2] https://github.com/nigno17/Synthetic-video-generator.

Fig. 2 Frames extracted from the videos generated by our synthetic video generator

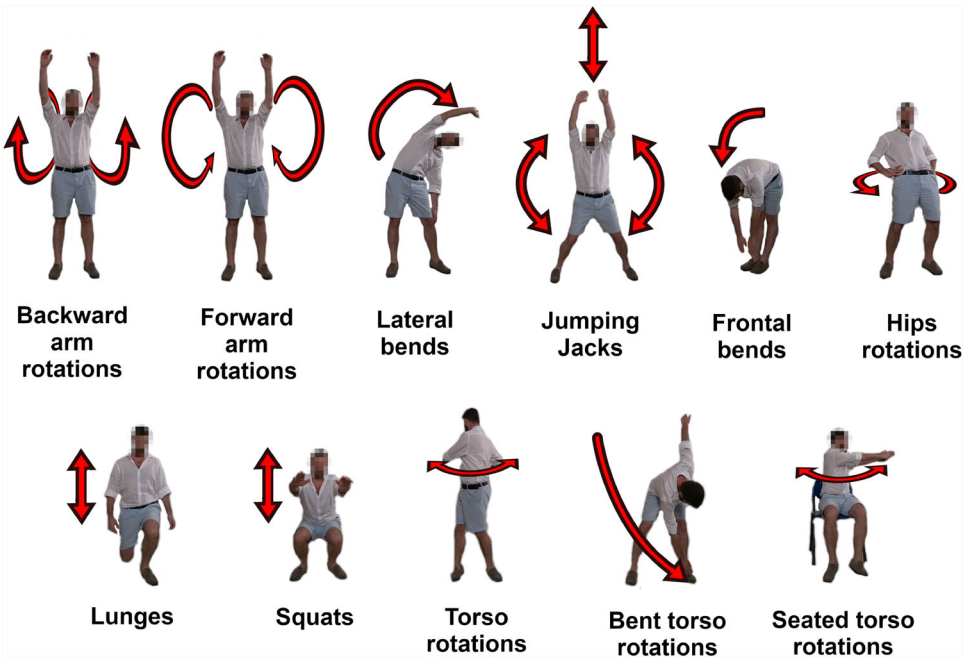

Fig. 3 Eleven exercises selected for our studies

4.1 The dataset

We first defined a suitable set of gentle gymnastic exercises to incorporate into the daily training routines of older individuals. To assist in this process, we enlisted the expertise of a professional personal trainer with 13 years of experience. Through collaboration with the personal trainer, we finalized a set of 11 exercises:

1. Backward arm rotations
2. Forward arm rotations
3. Lateral bends
4. Jumping jacks
5. Frontal bends
6. Hips rotations
7. Lunges
8. Squats
9. Torso rotations
10. Bent torso rotations
11. Seated torso rotations

Figure 3 depicts a visual representation of each of the 11 exercises.

To support our study, we initiated the collection of a small real-life video dataset in our laboratory, which we subsequently augmented by generating a larger synthetic dataset using our Synthetic Video Generator. We enlisted 23 subjects among students (14) and professors (9) from the Department of Mathematics and Computer Science at the University of Cagliari to participate in the collection of two distinct sets of videos. The initial set was recorded indoors utilizing the Intel RealSense D455 camera [54], while the

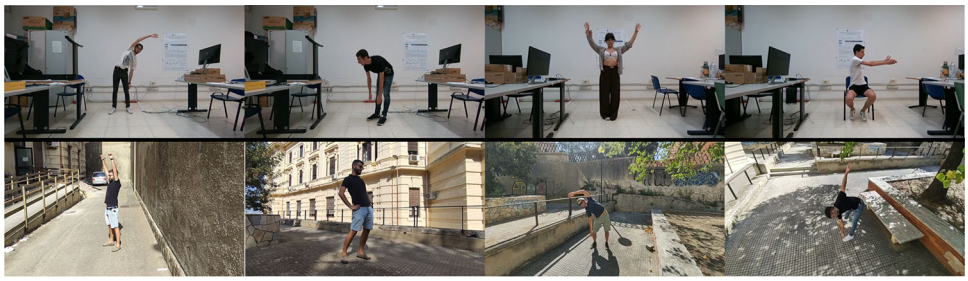

Fig. 4 Frames extracted from the collected videos. Upper row: frames from the training/validation set. Bottom row: frames from the test set

subsequent set employed the onboard camera of the Xiaomi Poco3 cellphone. Each subject executed each of the 11 exercises multiple times, resulting in a video captured for each repetition. The first twenty subjects underwent recording in a controlled environment within our laboratory. The position of the RealSense camera remained fixed in front of the subjects and remained unchanged throughout the entire recording session. Subsequently, this dataset was partitioned to form a training set comprising recordings from the first 15 subjects and a validation set containing recordings from the remaining five subjects. Refer to the upper row of Fig. 4 for exemplar frames extracted from the training and validation set videos. The recordings of the final three subjects were designated to compose our test set. These videos were captured outdoors, within the courtyard of our department, under varying light conditions, across multiple locations, and with diverse camera positions. This dataset aimed to replicate a more realistic setting. The bottom row of Fig. 3 showcases exemplar frames extracted from this dataset.

To summarize, we divided our data into three datasets:

1. *Training set:* This dataset, used for training, contains 1647 videos. Each video corresponds to one repetition of one of the 11 exercises executed by one of the first 15 students recorded in the laboratory. Each exercise was repeated roughly 10 times by each student.
2. *Validation set:* This dataset, used for validation, contains 539 videos. Each video corresponds to one repetition of one of the 11 exercises executed by the remaining five students recorded in the laboratory. Each exercise was repeated roughly 10 times by each student.
3. *Test set:* This dataset, used for testing, contains 329 videos. Each video corresponds to one repetition of one of the 11 exercises executed by one of the three students recorded outdoors. Each exercise was repeated roughly 10 times by each student.

Training, validation, and test collected datasets are accessible online through the provided link in the footnote.[3]

[3] https://drive.google.com/drive/folders/1GzkfOD9byPzOhIPMceiBB6DRdCHRX4Ud?usp=drive_link.

To train the Video Action Recognition module of our robotic coach, we augmented the collected training set generating a synthetic video set of data. We extracted the body motions from all the videos in the collected training set using the Mediapipe pose landmark detection model. We then imported those body motions as Unity animation clips into our synthetic video generator via a C# script. We generated a dataset of 5000 videos displaying a human avatar performing one of the 11 exercises based on the imported animations. Each video was generated by randomizing avatar appearance (age, ethnicity, height, gender, body mass, and clothes), background images, scene illumination (color, intensity, and direction), animation speed, and camera position and orientation. Figure 2 displays some example of generated frames.

The generated data was used to create a final "Augmented training set" that contains 6647 videos. It combines the 5000 videos generated with the synthetic video generator with the 1647 training videos collected in the laboratory.

5 Architecture of the system

In this section, we will analyze the final robotic coach architecture and the implementation of each of its modules in more detail (see Fig. 5). The following subsections describe the three main modules of the system one by one.

5.1 Video action recognition module

The role of the video action recognition module is to analyze video clips received from a camera placed next to the robot, showing elderly users performing the exercises. The module can recognize which of the 11 actions has been performed by the elderly person and determine if it is the requested exercise or not. The core of the module can be implemented using a deep learning video action recognition model.

Following the evaluation of our augmented dataset on two cutting-edge models for video action recognition (one based on convolutional neural networks (CNNs), known as I3D [55], and the other utilizing vision transformer (ViT) architecture [25], referred to as Timesformer [25]), we decided to implement the video action recognition module using the Timesformer model. ViT architectures are rapidly

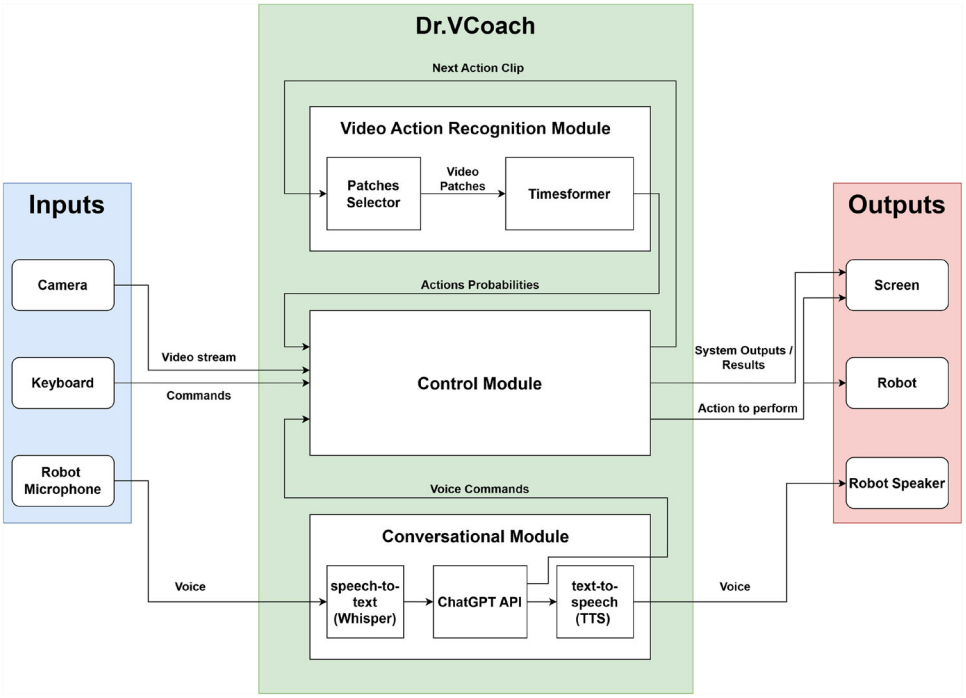

Fig. 5 Flow diagram of interactions between modules

becoming the de facto standard for video and image processing and classification. The Timesformer architecture is an extension of the ViT model tailored for video action recognition. The attention model in Timesformer not only encompasses the spatial domain, but also extends its influence over the temporal domain. The input to the Timesformer model consists of a sequence of RGB frames, with each frame divided into N non-overlapping patches. The self-attention mechanism, known as divided space-time attention, initially calculates the temporal attention between a patch and all corresponding spatial patches in other frames. Subsequently, it computes the spatial attention for the resulting temporal encoding with respect to all other patches within the analyzed frame.

More specifically, the video action recognition module receives the pre-cropped video clip of the next action to analyze from the Control Module. The clip is preprocessed, and each frame is divided into patches that are linearly transformed into embedding vectors. These embedded patches are then analyzed by the Timesformer model using the divided space-time attention mechanism. For each video clip, the Timesformer produces a probability distribution across 11 possible actions. These probabilities are then sent back to the control module for further analysis.

5.2 Conversational module

The conversational module serves as an interface to manage verbal interactions between the elderly user utilizing the system and the robotic coach. It should be capable of engaging in conversation with the user, understanding specific verbal commands, and providing vocal updates on the status of the training session.

We have implemented our conversational module using the ChatGPT API. More specifically, we leverage the API's innovative "function calling" capability. This feature enables new ChatGPT models to identify specific questions or commands from users and respond with a structured call request instead of standard text replies. By incorporating function calling, chatbots gain the ability to interact with other systems, allowing the GPT models to handle inquiries that require the invocation of an external function. Put simply, function calling offers an additional means of instructing AI models on how to engage with the external environment.

We designed the conversational module as a chatbot using the python ChatGPT API. We used the function calling capability to define the following function calls:

1. **Function:** "start_session". **Description:** "Start the daily training section for user". **Effect:** starting the Control Module for a specific user.
2. **Function:** "stop_session". **Description:** "Stop the training section". **Effect:** stopping the Control Module.
3. **Function:** "start_exercise". **Description:** "Start monitoring the next exercise". **Effect:** the Control Module start the monitoring phase for the next exercise.
4. **Function:** "repeat_exercise". **Description:** "Repeat monitoring the last exercise". **Effect:** the Control Module start the monitoring phase for the past exercise again.

5. **Function:** "show_exercise". **Description:** "Show me the exercise again". **Effect:** showing again the exercise via video and performed by the robot.

To enable voice control of the conversational module and hear its responses, we utilized the capabilities of OpenAI's speech-to-text (Whisper) and text-to-speech (TTS) models. We developed a submodule connected with the conversational module that is capable of:

1. Receiving all responses from the conversational module in real-time as strings of text.
2. Generating audio tracks for each received string via the TTS model and playing them.
3. Listening to the user for verbal commands or conversational responses via the laptop microphone.
4. Converting audio tracks received from the microphone into strings of text to be passed to the conversational module using the Whisper model.

5.3 Control module

The control module is the core of our robotic coach and synchronizes all the other modules. Specifically, the functions of the control module are:

1. Initialize a training session receiving a command via the conversational module.
2. Select the sequence of exercises for the daily training session based on the schedule defined by a therapist.
3. Show the exercises to the user (with verbal descriptions) both via a video on a screen and performed by the robot through the control module.
4. Monitor the execution of each exercise via the video action recognition module.
5. Show the evaluation of the training session to the user/patient.

Before integrating all the blocks of our system, we decided to implement a prototype of the control module that runs on a laptop and does not interact with the robot and the conversational module. The prototype is controlled via keyboard and mouse. Figure 6 shows four different screenshots of the prototype running. The program can store info and data of multiple patients. At first, the prototype asks the patient to select an existing account or create a new one (Fig. 6A). After logging in, a screen with the daily training list of exercises is shown. In this screen, the patient can play a sample video and read a verbal explanation of each exercise (Fig. 6B). When the patient is ready, the monitoring process for each exercise starts with the recording screen. In this modality, the patient must copy the exercise shown by a video on the screen for several repetitions. Other than the sample video, the real-time video of the patients captured from the webcam and the name of the exercise that has been executed are shown on screen (Fig. 6C). During the monitoring, the videos captured from the camera for each repetition are stored and analyzed by the video action recognition module (the Timesformer model trained on our augmented dataset). After the execution of each exercise, the program shows a screen with the evaluation of the execution (Fig. 6D). In the summary page, for each repetition, it is possible to see:

1. The recorded video.
2. The ordered list of the prediction confidence of our action recognition model for each exercise.
3. An icon showing the quality of the execution:
   (a) Green circle—good execution (the tested exercise is predicted as the more probable)
   (b) Yellow circle—average execution (the tested exercise is predicted as the second or third more probable)
   (c) Red circle—bad execution (the tested exercise is out of the top three predictions)

Finished the training session, the program stores all the video and evaluation results in a separated folder for each registered patient. The working code of the control module prototype is available online at the provided link in the footnote.[4]

Before implementing the final system, we generated robot animations for the exercises on the Zora robot using the Softbank Choregraphe Suite. However, due to the Zora robot's limited motion capabilities, certain exercises could not be faithfully animated. Specifically, the Jumping Jacks exercise could not be implemented as the Zora robot is unable to perform jumping motions.

6 System at work

We decided to conduct two different tests to evaluate the proposed robotic coach system. First, we tested the validity of the augmented dataset generated using our synthetic video generator. This step was essential before using the augmented dataset to train the Timesformer model in the video action recognition module of the final system. In the second test, we integrated all the system modules and tested the resulting final version of the robotic coach on nine healthy adult subjects aged over 50. After the test, we asked the subjects to complete a System Usability Scale (SUS) survey and analyzed their responses.

[4] https://github.com/nigno17/DrVCoach.

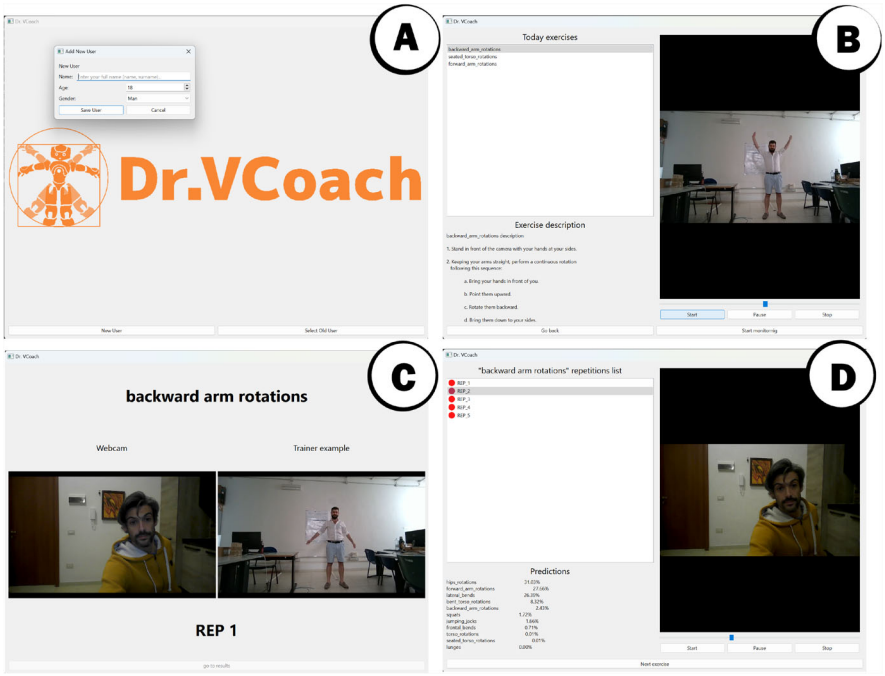

Fig. 6 Control Module in action: **A** Login screen with the "Add New User" window open. **B** Screen displaying the exercises scheduled for the day. **C** Monitoring and recording interface. **D** Results display screen

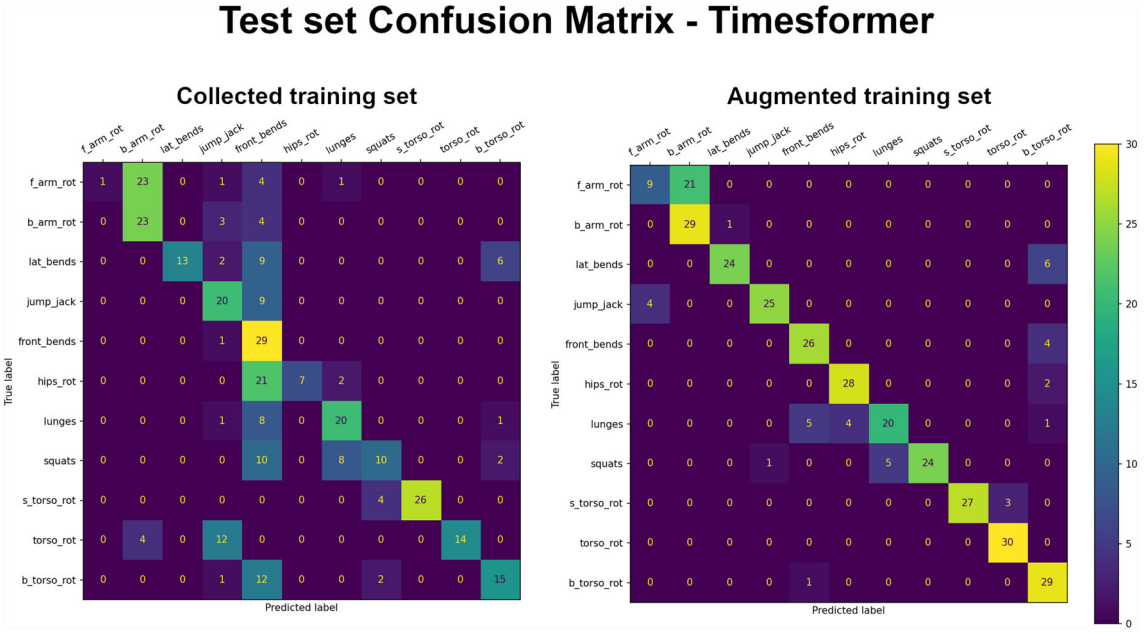

Fig. 7 Confusion Matrices of the Timesformer model trained on the collected and augmented dataset. The model trained on the augmented dataset has a high accuracy (the matrix is almost diagonal), presenting some confusion only between the backward and forward arm rotations, two exercises visually very similar

6.1 Dataset validation

We decided to evaluate our datasets on two cutting-edge models: one based on convolutional neural networks (CNNs), known as I3D, and the other utilizing vision transformer (ViT) architecture, referred to as Timesformer. The experiments were conducted on a laptop featuring an AMD Ryzen 9 5900HX CPU and an NVIDIA GeForce RTX 3080 Laptop GPU. The models, which were pretrained on ImageNet, were trained for 15 epochs in each experiment. Video frames were processed by resizing the smaller dimension to 256 pixels and applying a central crop to produce 224x224 images. After cropping, ImageNet normalization was applied to standardize the frames. The training process for both networks

employed the stochastic gradient descent (SGD) optimizer, configured with a learning rate of 0.005, a momentum of 0.9, and a weight decay of 0.0001.

The two-stream inflated 3D ConvNet (I3D) is a video action recognition model based on 3D CNNs. The Timesformer architecture is an extension of the ViT model tailored for video action recognition.

The two models have been trained on both our collected and augmented dataset. The I3D obtained an accuracy on our test set of 19% trained on the collected and 83% trained on the augmented. The Timesformer had a better accuracy when trained on the collected dataset, 54%, and an accuracy in line with I3D when trained on the augmented dataset, 82%. In general, the Timesformer showed a better generalization (better results trained on the small and less varied collected dataset), faster training time during fine-tuning, and good results trained on the varied augmented dataset (over 80% accuracy, confusion matrix depicted in Fig. 7). Moreover, ViT architectures are rapidly becoming the de facto standard for video and image processing and classification. For these reasons, we decided to implement the Dr. VCoach "Video Action Recognition Module" using the Timesformer model trained on our augmented dataset.

6.2 Final system validation

The final setup of the Dr. VCoach system consists of the Zora robot and a laptop running all the modules, placed in front of the user/patient. The patient interacts via voice with the conversational module's chatbot to control the system. In this way, the system can provide more information about the exercises to the user, both through videos and the robot's embodied animation.

To perform the final test and validate the robotic coach, all the components presented in Sect. 5 were integrated together. First, all the functions of the conversational module were connected to the control module prototype. This allowed the execution of connected commands on the control module when a patient gives one of the specified commands via function calling. (For example, when the patient asks to start the daily training session, the control module program is launched.) Subsequently, the robot was integrated into the system. Whenever the control module displays an exercise video to the patient, a signal is sent to the robot to start the corresponding animation. However, for the Jumping Jack exercise, which cannot be performed by the robot, only the video is played.

We selected nine healthy adults over 50 years old to validate the final system. A unique ID was assigned to each subject to maintain anonymity. The subjects' ages range from 51 to 86 years old, with an equal distribution between men (5 subjects) and women (4 subjects). Table 2 presents a list of all subjects with their IDs, ages, and genders.

Table 2 List of the selected subjects for the test

Subjects list		
Subject id	Age	Gender
Sub1	51	Man
Sub2	56	Woman
Sub3	86	Man
Sub4	54	Man
Sub5	82	Woman
Sub6	77	Man
Sub7	69	Woman
Sub8	54	Man
Sub9	57	Woman

After an initial explanation of how the system works, each subject used the system for three training sessions. During each session, the robotic coach presented the subject with three different exercises randomly selected from the pool of 11 exercises defined previously. The subject then performed each exercise for 5 repetitions, after which the control module displayed the evaluation screen. Each exercise could be repeated only twice during the three training sessions. Consequently, each subject performed a minimum of five and a maximum of nine different exercises during the test. Figure 8 illustrates the distribution of training sessions for each exercise throughout the entire test phase. On average, each exercise was presented an equal number of times across all sessions, with the exception of lunges, which unfortunately appeared less frequently than the others. All the subjects signed an informed consent form before commencing the training sessions. The form contains information regarding the purpose, process, potential benefits, and risks of the study, as well as details on data collection procedures, time commitment, voluntary participation, and the right to withdraw (without prejudice to care). Additionally, the form provides assurance of confidentiality, stating that research records identifying participants will be kept confidential to the extent permitted by applicable laws and regulations.

The system stores the classification predictions and the captured videos of all the sessions, with each subject having their own folder. To evaluate the effectiveness of the training coach, we analyzed the top-1 and top-3 accuracy of the Action Video Recognition module. Figure 9 illustrates the accuracy of the system across all training sessions and grouped by exercise. Overall, the system's accuracy is slightly lower than the results obtained in our previous test set, with the majority of exercises achieving a top-1 accuracy of over 70% and top-3 accuracy of over 90%. The challenge of the trained model in distinguishing between forward and backward arm rotations is also evident in this final test. Specifically, the top-1 accuracy for these two exercises is low (45% and 68%, respectively), while their top-3 accuracy is

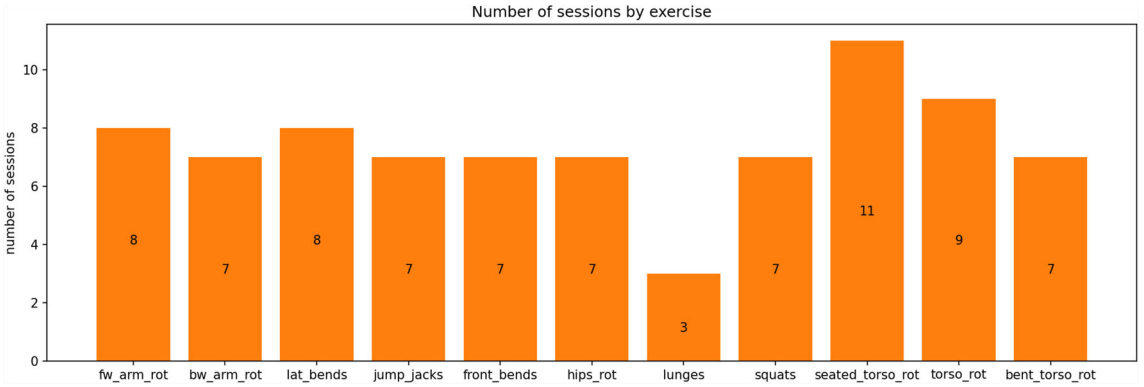

Fig. 8 Number of total training sessions on each exercise for all the subjects

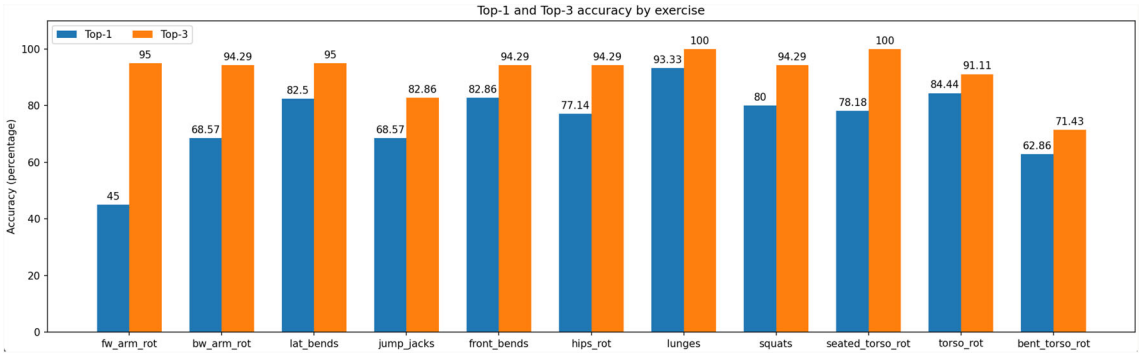

Fig. 9 Top-1 and Top-3 accuracy obtained by the subjects for each exercise during their training sessions

close to 100% (95% and 94%). These results are noteworthy for two reasons:

1. The age of the subjects in this study is higher than that of the subjects enrolled in collecting our previous test set (who were in their 20 s), potentially resulting in poorer exercise performance.
2. The training sessions were recorded in participants' homes, in an uncontrolled environment. This led to high variance in illumination conditions, distracting elements in the environment (such as other individuals not performing the exercises), and instances of body parts being cropped (refer to Fig. 13 for examples).

Another important factor in evaluating the effectiveness of the system is its ability to recognize when a specific exercise is performed poorly. In such cases, the system should not classify the correct exercise as the most probable one. To gain insight into this aspect, we analyzed the top-1 and top-3 accuracy achieved by each subject during all training sessions (results are depicted in Fig. 10). We anticipated a decrease in accuracy with increasing age of the subjects, as older individuals are expected to perform the exercises less effectively compared to younger ones. Indeed, the two subjects over 80 years old achieved significantly lower accuracy than the other subjects (top-1 of 33% and 44%), followed closely by subject 6 (77 years old, top-1 accuracy of 66%).

After the training sessions, we asked the subjects to complete a survey regarding their experience. The survey was divided into four sections:

1. Background: Subjects were asked to provide their personal information, including name, date of birth, gender, and email address.
2. System Usability Scale (SUS): This section consisted of a set of standard questions regarding the usability of the system. Each question was rated on a scale from 1 (strongly disagree) to 5 (strongly agree). Figure 11 displays the list of questions and the average ratings obtained across all subjects.
3. DrVCoach results analytics: This section contained questions regarding the quality of the training sessions. Each question was rated on a scale from 1 (low) to 5 (high). Figure 12 presents the list of questions and the average ratings obtained across all subjects.
4. Open questions: Subjects were asked to identify the strengths and weaknesses of the system, as well as suggest any additional features to be included.

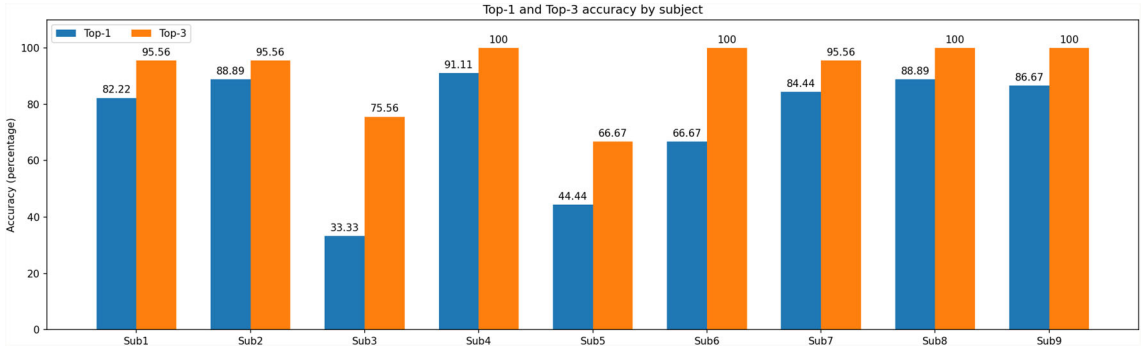

Fig. 10 Top-1 and Top-3 accuracy obtained by each subject during his/her training sessions

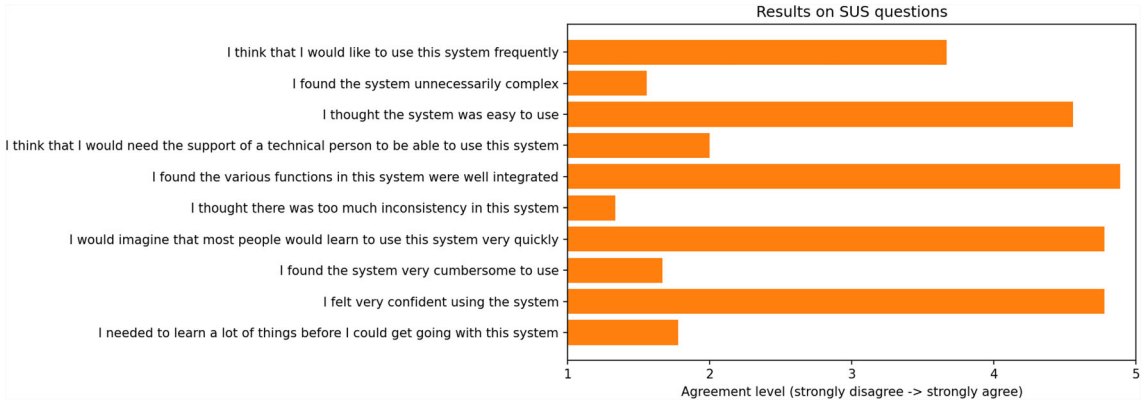

Fig. 11 System usability scale (SUS) survey's questions with the average answer rating across the subjects

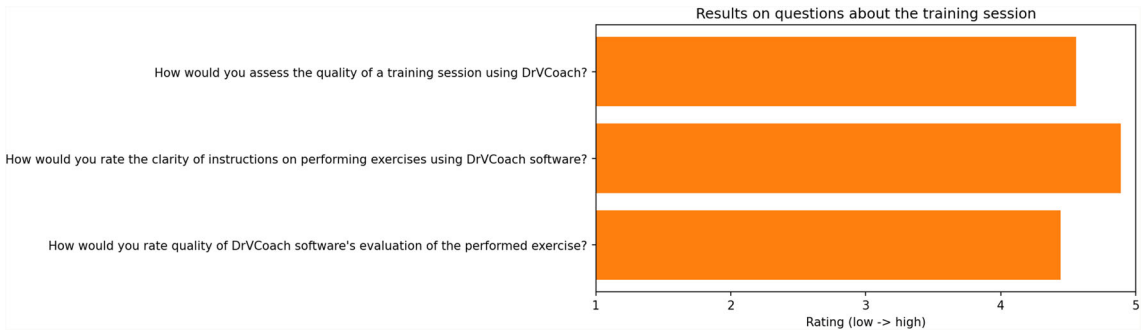

Fig. 12 List of survey's questions on the training session with the average answer rating across the subjects

The survey results indicate a positive response from the subjects for both the SUS questions and those regarding the quality of the training sessions. Subjects found the system easy to use and well-integrated; however, they expressed some uncertainty about using it frequently in their everyday lives.

When asked about the strengths of the system, most of the subjects highlighted its ability to help them maintain their physical fitness ("it helps older people maintain their physical fitness," "perfect for home training," and "it helps me move my body"), as well as the ease of use of the system ("the system presents the exercises clearly," "it is easy to learn and very useful," and "the exercise explanations are clear enough"). Regarding the weaknesses of the system, some subjects did not identify any, while others noted the need to familiarize themselves with the program's functionalities before use ("I need to familiarize myself with the system's functionalities before using it," "the system lacks a user manual or tutorial"), and the absence of a warm-up session ("I believe it could include a warm-up session at the beginning").

Only three subjects mentioned additional features they would like to see: "the system should include a warm-up phase before starting the training session," "I would appreciate having more exercises," and "I would like an audio

Fig. 13 Frames extracted from the videos captured during the validation test of the system

explanation of each exercise in addition to the practical examples."

7 Conclusion

In this paper, we introduced a robotic coach capable of guiding elderly patients through their daily physical exercise routines to maintain an active lifestyle and age healthily. The robotic coach can verbally interact with patients, initiate training sessions, describe and demonstrate exercises to be performed, monitor and evaluate exercise execution using an external camera, and display the subjects' performances on a screen. Additionally, we provided a collected dataset of human subjects performing exercises in front of the camera and developed a tool to augment this dataset by generating randomized synthetic videos. The entire system was tested on nine elderly subjects in their homes, followed by the administration of a SUS survey to each subject.

Thanks to our generative data augmentation architecture and the video action recognition module based on vision transformer, the robotic coach successfully recognized and evaluated the execution of exercises performed by the subjects during testing. The system demonstrated a high level of robustness in uncontrolled environments where training sessions were recorded, including variations in illumination conditions, distracting elements in the environment, and instances of body parts being cropped.

Upon analyzing the subjects' responses to the survey, we found a high level of acceptance of the robotic coach for guiding and motivating them during training. When asked about the system's strengths, subjects highlighted its usefulness and ease of use. The only drawbacks mentioned were the time required to become familiar with the system and the absence of a warming-up stage before starting the training session. The positive reception from the nine elderly subjects during testing is encouraging for the broader adoption of robotic coaches and assistive robots in healthcare scenarios such as physical training and rehabilitation.

We plan to extend our work in the future to address some of the current limitations and explore additional avenues for enhancement. The first step will involve improving the action recognition module by upgrading the system to handle multiple patients on screen and recognizing actions that involve human–human and human–object interactions. To achieve these enhancements, our unity-based synthetic generator must be updated to generate videos featuring multiple actors, human interactions, and the inclusion of handheld and interactive objects.

The second step will focus on integrating smart healthcare features into the system. Specifically, we plan to add a cloud database to store patient data, training programs, and progress. Additionally, we intend to develop a mobile app connected to the database, enabling both patients and therapists to interact with the system. These features will be crucial for therapists, allowing them to remotely monitor patient progress, update training programs, and provide real-time feedback.

Finally, we aim to test the completed system in a real-world scenario, with patients using the system over an extended period (e.g., 1 month). During this final test, therapists will select appropriate training programs for each patient and closely monitor their progress.

Supplementary information

If your article has accompanying supplementary file/s, please state so here.

Authors reporting data from electrophoretic gels and blots should supply the full unprocessed scans for key as part of their Supplementary information. This may be requested by the editorial team/s if it is missing.

Please refer to Journal-level guidance for any specific requirements.

Acknowledgements This research was funded by the European Union's Horizon 2020 Marie Skłodowska-Curie Actions Individual Fellowships under the Grant No. 101031646.

Author Contributions Nino Cauli was responsible for the software implementation, the design and execution of the experiments, and the validation of the results. He also drafted the initial version of the manuscript. Diego Reforgiato Recupero contributed through supervision of the work, project administration and critical revision of the manuscript.

Funding Open access funding provided by Università degli Studi di Cagliari within the CRUI-CARE Agreement. This research was funded by the European Union's Horizon 2020 Marie Skłodowska-Curie Actions Individual Fellowships under the Grant No. 101031646.

Data availability Training, validation, and test collected datasets are accessible online through the following link: https://drive.google.com/drive/folders/1GzkfOD9byPzOhIPMceiBB6DRdCHRX4Ud?usp=drive_link.

Declarations

Ethics approval and consent to participate Informed consent was obtained from all subjects involved in the study.

Consent for publication Written informed consent has been obtained from the patients to publish this paper.

Code availability The synthetic video generator and the MediaPipe scripts for extracting skeletal animations from videos and converting them into Unity animation clips are accessible online through the following link: https://github.com/nigno17/Synthetic-video-generator. The working code of the control module prototype is available online at the following link: https://github.com/nigno17/DrVCoach.

References

1. Mukherjee D, Gupta K, Chang LH, Najjaran H (2022) A survey of robot learning strategies for human-robot collaboration in industrial settings. Robot Comput-Integrat Manuf 73:102231. https://doi.org/10.1016/j.rcim.2021.102231
2. Aly AAI, Abbasimoshaei A, Kern TA (2022) Developing a VR training environment for fingers rehabilitation. Springer. https://doi.org/10.15480/882.4368 . http://hdl.handle.net/11420/12817
3. Tsiakas K, Papakostas M, Theofanidis M, Bell M, Mihalcea R, Wang S, Burzo M, Makedon F (2017) An interactive multisensing framework for personalized human robot collaboration and assistive training using reinforcement learning. In: Proceedings of the 10th international conference on PErvasive technologies related to assistive environments. PETRA '17, Association for Computing Machinery, New York, pp 423–427. https://doi.org/10.1145/3056540.3076191
4. Lee MH, Siewiorek DP, Smailagic A, Bernardino A, Badia SBi (2019) Learning to assess the quality of stroke rehabilitation exercises. In: Proceedings of the 24th international conference on intelligent user interfaces, pp 218–228
5. Lee MH, Siewiorek DP, Smailagic A, Bernardino A, Badia SB (2020) Towards personalized interaction and corrective feedback of a socially assistive robot for post-stroke rehabilitation therapy. In: 2020 29th IEEE international conference on robot and human interactive communication (RO-MAN), IEEE, pp 1366–1373
6. Cauli N, Massa SM, Recupero DR, Riboni D (2023) Sensor datasets for human daily safety and well-being. Enabling person-centric healthcare using ambient assistive technology: personalized and patient-centric healthcare services in AAT. Springer, Cham, pp 1–26
7. Mostajeran F, Steinicke F, Ariza Nunez OJ, Gatsios D, Fotiadis D (2020) Augmented reality for older adults: exploring acceptability of virtual coaches for home-based balance training in an aging population. In: Proceedings of the 2020 CHI conference on human factors in computing systems, pp 1–12
8. Beer JM, Smarr C-A, Chen TL, Prakash A, Mitzner TL, Kemp CC, Rogers WA (2012) The domesticated robot: design guidelines for assisting older adults to age in place. In: Proceedings of the seventh annual ACM/IEEE international conference on human–robot interaction, pp 335–342
9. Esposito R, Fiorini L, Limosani R, Bonaccorsi M, Manzi A, Cavallo F, Dario P (2016) Supporting active and healthy aging with advanced robotics integrated in smart environment. Optimizing assistive technologies for aging populations. IGI Global, Hershey, pp 46–77
10. Winkle K, Caleb-Solly P, Turton A, Bremner P (2018) Social robots for engagement in rehabilitative therapies: Design implications from a study with therapists. In: Proceedings of the 2018 ACM/IEEE international conference on human–robot interaction, pp 289–297
11. Čaić M, Avelino J, Mahr D, Odekerken-Schröder G, Bernardino A (2020) Robotic versus human coaches for active aging: an automated social presence perspective. Int J Soc Robot 12(4):867–882
12. Loos HM, Reinkensmeyer DJ, Guglielmelli E (2016) Rehabilitation and health care robotics. Springer handbook of robotics, pp 1685–1728
13. Tapus A, Ţăpuş C, Matarić MJ (2008) User-robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy. Intel Serv Robot 1:169–183
14. Gockley R, MatariĆ MJ (2006) Encouraging physical therapy compliance with a hands-off mobile robot. In: Proceedings of the 1st ACM SIGCHI/SIGART conference on human–robot interaction, pp 150–155
15. Matarić MJ, Eriksson J, Feil-Seifer DJ, Winstein CJ (2007) Socially assistive robotics for post-stroke rehabilitation. J Neuroeng Rehabil 4:1–9
16. Fasola J, Matarić MJ (2013) A socially assistive robot exercise coach for the elderly. J Human-Robot Interact 2(2):3–32
17. Swift-Spong K, Short E, Wade E, Matarić MJ (2015) Effects of comparative feedback from a socially assistive robot on self-efficacy in post-stroke rehabilitation. In: 2015 IEEE international conference on rehabilitation robotics (ICORR), IEEE, pp 764–769
18. Nguyen SM, Tanguy P, Remy-Neris O (2016) Computational architecture of a robot coach for physical exercises in kinaesthetic rehabilitation. In: 2016 25th IEEE international symposium on robot and human interactive communication (RO-MAN), pp 1138–1143
19. Lee MH, Siewiorek DP, Smailagic A, Bernardino A, Badia SB (2024) Enabling AI and robotic coaches for physical rehabilitation

therapy: iterative design and evaluation with therapists and post-stroke survivors. Int J Soc Robot 16(1):1–22
20. Ji S, Xu W, Yang M, Yu K (2012) 3d convolutional neural networks for human action recognition. IEEE Trans Pattern Anal Mach Intell 35(1):221–231
21. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems, pp 568–576
22. Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4694–4702
23. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv, Neural Inf Process Syst, p 30
24. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al (2020) An image is worth 16x16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929
25. Bertasius G, Wang H, Torresani L (2021) Is space-time attention all you need for video understanding? In: ICML, vol. 2, p. 4
26. Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C (2021) Vivit: a video vision transformer. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 6836–6846
27. Jaegle A, Gimeno F, Brock A, Vinyals O, Zisserman A, Carreira J (2021) Perceiver: general perception with iterative attention. In: International conference on machine learning, PMLR, pp 4651–4664
28. Selva J, Johansen AS, Escalera S, Nasrollahi K, Moeslund TB, Clapés A (2023) Video transformers: a survey. IEEE Trans Patt Anal Mach Intell 45(11):12922–12943
29. Lee N, Choi W, Vernaza P, Choy CB, Torr PH, Chandraker M (2017) Desire: distant future prediction in dynamic scenes with interacting agents. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 336–345
30. Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data 6(1):1–48
31. Khalifa NE, Loey M, Mirjalili S (2021) A comprehensive survey of recent trends in deep learning for digital images augmentation. Artif Intell Rev 55(3):2351–2377
32. Tremblay J, To T, Sundaralingam B, Xiang Y, Fox D, Birchfield S (2018) Deep object pose estimation for semantic robotic grasping of household objects. arXiv preprint arXiv:1809.10790
33. Tobin J, Fong R, Ray A, Schneider J, Zaremba W, Abbeel P (2017) Domain randomization for transferring deep neural networks from simulation to the real world. In: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE, pp 23–30
34. Cauli N, Reforgiato Recupero D (2022) Survey on videos data augmentation for deep learning models. Fut Internet 14(3):93
35. Marszałek M, Laptev I, Schmid C (2009) Actions in context. In: IEEE conference on computer vision & pattern recognition
36. Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) HMDB: A large video database for human motion recognition. In: Proceedings of the international conference on computer vision (ICCV)
37. Soomro K, Zamir AR, Shah M (2012) Ucf101: a dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402
38. Smaira L, Carreira J, Noland E, Clancy E, Wu A, Zisserman A (2020) A short note on the kinetics-700-2020 human action dataset. arXiv preprint arXiv:2010.10864
39. Rodriguez MD, Ahmed J, Shah M (2008) Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: 2008 IEEE conference on computer vision and pattern recognition, IEEE, pp 1–8
40. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: CVPR
41. Singh S, Velastin SA, Ragheb H (2010) Muhavi: a multicamera human action video dataset for the evaluation of action recognition methods. In: Advanced video and signal based surveillance (AVSS), 2010 seventh IEEE international conference on, IEEE, pp 48–55
42. Li W, Mahadevan V, Vasconcelos N (2013) Anomaly detection and localization in crowded scenes. IEEE Trans Patt Anal Mach Intell 36(1):18–32
43. Weinland D, Ronfard R, Boyer E (2006) Free viewpoint action recognition using motion history volumes. Comput Vis Image Underst 104(2–3):249–257
44. To T, Tremblay J, McKay D, Yamaguchi Y, Leung K, Balanon A, Cheng J, Hodge W, Birchfield S (2018) NDDS: NVIDIA deep learning dataset synthesizer. https://github.com/NVIDIA/Dataset_Synthesizer
45. Hwang H, Jang C, Park G, Cho J, Kim I (2021) Eldersim: a synthetic data generation platform for human action recognition in eldercare applications. IEEE Access 11:9279–9294
46. robotics Z. Zora/Nao robot hompage. https://www.zorarobotics.be/robots/nao. Accessed 05 Oct 2022
47. SoftBank Robotics: NAO robot. https://www.softbankrobotics.com/emea/en/nao. Accessed 16 Sept 2021
48. robotics A. Choregraphe documentation page. http://doc.aldebaran.com/2-4/software/choregraphe/index.html. Accessed 27 Oct 2023
49. Technologies U. Unity homepage. https://unity.com/. Accessed 05 Oct 2022
50. Technologies U. Unity Perception Package github page. https://github.com/Unity-Technologies/com.unity.perception. Accessed 10 Oct 2023
51. Technologies U. Unity Synthetic Humans github page. https://github.com/Unity-Technologies/com.unity.cv.synthetichumans. Accessed 10 Oct 2023
52. Google: Mediapipe homepage. https://developers.google.com/mediapipe. Accessed 10 Oct 2023
53. Bazarevsky V (2020) Blazepose: on-device real-time body pose tracking. arXiv preprint arXiv:2006.10204
54. Intel: RealSense homepage. https://www.intel.com/content/www/us/en/architecture-and-technology/realsense-overview.html. Accessed 05 Oct 2022
55. Carreira J, Zisserman A (2017) Quo vadis, action recognition? A new model and the kinetics dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 6299–6308