Human pose estimation based in-home lower body rehabilitation system

Ying Li
Department of Computer Science
University of Massachusetts Lowell
Lowell, USA
ying_li@student.uml.edu

Chenxi Wang
Department of Electrical and Computer Engineering
University of Massachusetts Lowell
Lowell, USA
chenxi_wang1@student.uml.edu

Yu Cao
Department of Computer Science
University of Massachusetts Lowell
Lowell, USA
yu_cao@uml.edu

Benyuan Liu
Department of Computer Science
University of Massachusetts Lowell
Lowell, USA
benyuan_liu@uml.edu

Joanna Tan
Physical Therapy Department
Encompass Rehabilitation Hospital
Woburn, USA
fangtan777@gmail.com

Yan Luo
Department of Electrical and Computer Engineering
University of Massachusetts Lowell
Lowell, USA
yan_luo@uml.edu

***Abstract*—In this paper, we design, develop and evaluate an in-home lower body rehabilitation system based on a novel lightweight human pose estimation model. To achieve that, we first create a lower body rehabilitation dataset of 500,000 images with each image annotated with the ground truth joint point locations. The dataset consists of 31 different types of lower body rehabilitation activities from twenty volunteers. After that, we design a lightweight but powerful neural network model, which runs on a smartphone, to estimate human pose. Furthermore, we develop a series of principles for evaluating in-home rehabilitation activities of patients in terms of the range of motion and duration of activities. For the concern of privacy, all the data collected from patients are encrypted, stored and processed locally on patients' own smartphones. Only the sanitized evaluation reports are uploaded and shared with the patients' primary doctors. Our model achieves 70.8 in AP score on the COCO val2017 set with only 4.7M parameters and 1.0 GFLOPs. Using our system, patients can perform lower body rehabilitation activities at home and obtain evaluation report without the presence of physical therapists. We believe our system can greatly facilitate in-home rehabilitation and reduce the cost for patients.**

***Index Terms*—in-home lower body rehabilitation system, human pose estimation**

I. Introduction

Nowadays, physical therapies play a critical role in rehabilitation for post-operation patients and patients with a wide variety of diseases. However, physical therapies can be very expensive and inconvenient. A physical therapy session usually costs from $50 to $350 or more in the U.S. and a complete physical therapy process typically requires 2 to 3 sessions at the clinical centers per week for months and even years. As a result, physical therapies can be unaffordable for many patients especially for those not covered by medical insurance. To reduce in-clinic visits, in-home computer-assisted physical therapy solutions have attracted extensive attention. However, in-home rehabilitation can still be costly because it requires the presence of therapists for supervision and evaluation. Another important challenge is that the in-home environment is often computing resource constrained and thus calls for a computationally inexpensive system such as a mobile smartphone. Motivated by these observations, the goal of our research is to make rehabilitation at home both affordable and portable in a computing resource poor environment.

In addressing the challenges of human supervision and evaluation of physical activities, we focus on the intrinsic problem of a human pose estimation that detects the positions of human body key joint points (shoulder, elbow, knee, wrists, etc.) from a single image, a series of images by a single camera, or multiple images from multiple cameras. In this paper, we specifically aim at the scenario of using single-image from a single camera which can lead to low-cost and effective solutions. The techniques developed for this scenario will provide an important foundation and insights into more general multi-camera multi-image cases.

Due to the powerful representational capabilities of Convolutional Neural Networks (CNNs), the research on human pose estimation has witnessed significant advances recently [7] [8] [10] [11] [12] [13] [26]. For instance, High Resolution Network (HRNet) [13] leverages multiple resolution branches throughout the whole network and achieves the state-of-the-art performance on public datasets, such as COCO Keypoints Detection Dataset [14] and MPII [15]. However, these state-of-the-art solutions are too complicated to be deployed on mobile devices. For example, the number of parameters in HRNet model and Simple Baseline [12] model are up to 63 million and 68 million, respectively, rendering them impractical for a mobile device environment with limited computing resources. While it is possible to deploy the models as web services on powerful servers in the cloud, the privacy of patients is at risk as the original patient images are uploaded, which may violate strict privacy laws such as HIPAA and GDPR if not carefully handled. For this concern, we believe that a better solution is to have the system run locally on mobile devices.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

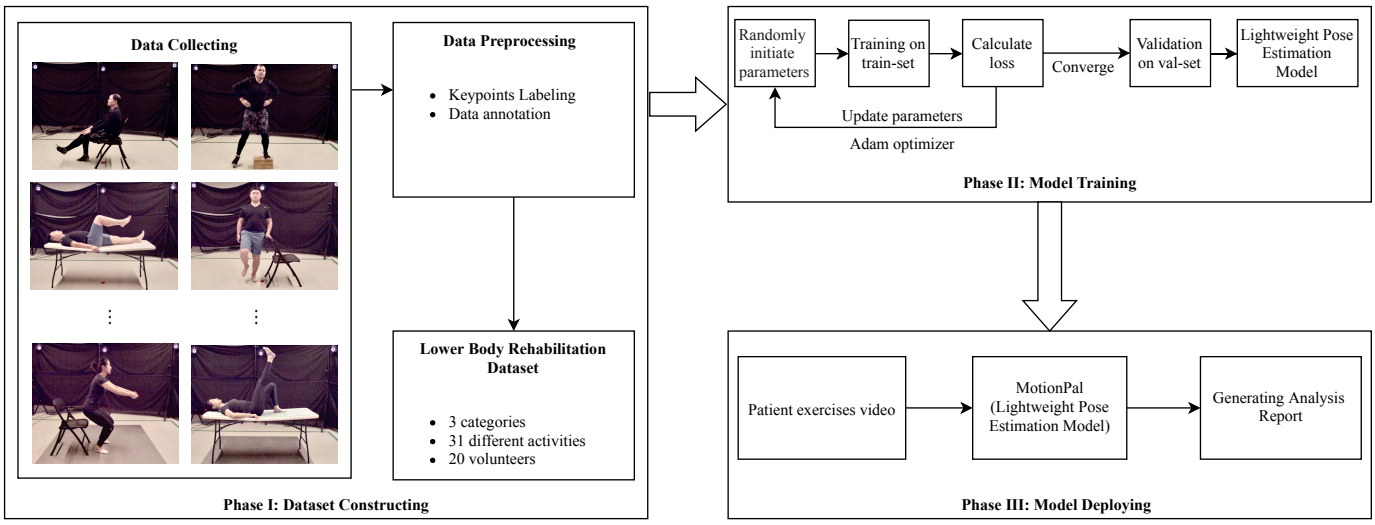


Fig. 1: Illustration of the three phases of our research. In Phase I, we collect and preprocess data to produce Lower Body Rehabilitation Dataset. Phase II shows our training process of the proposed Lightweight Pose Estimation Model. After that, we deploy it in the last phase.

In this paper, we propose a lightweight HRNet. In particular, we introduce depth-wise and dilated convolutional layers in the original model to significantly reduce both the parameter size and computation cost. To further improve the accuracy, we add attention branches in the proposed model with very little extra computation operations, which extract additional features by compressing and recovering the channels of the input feature map. With these modifications, our model can be deployed in smartphones and achieve high accuracy.

We also address in this paper the limitations of available datasets. Current public datasets for human pose estimation tasks such as Leeds Sports Poses (LSP) [1], MPII [15] and COCO Keypoints Detection [14], etc., are all designed for general purposes. Activities recorded in these datasets are very different from the rehabilitation activities. Moreover, we find that the ground truth locations of the body key joint points can be inaccurate in these datasets. For example, in COCO Keypoints Detection dataset, the shoulder location may be distributed across the whole shoulder part, which is unacceptable for the fine-grained activity evaluation and analysis required for rehabilitation assessment. To this end, we create a new dataset of lower body rehabilitation exercises using a 3D capture system. A total of 31 different types of rehabilitation activities from twenty volunteers are collected under the guidance of a physical therapist. After raw videos and annotations are collected, we review and refine the keypoint positions to ensure accurate annotations for each frame. To the best of our knowledge, this is the first rehabilitation activities key points detection and evaluation dataset. Experiments show that the models trained on our dataset can achieve excellent performance.

We implement our models in a prototype system, which consists of three major components: a mobile application (with a built-in lightweight HRNet model) for patients, a mobile application for doctors, and a web server to share data between patients and their primary doctors. The application for patients is the key part of our system, collecting images of rehabilitation activities by the patients and processing them locally. Only the final analysis reports are uploaded to the server and shared with doctors. After reviewing the reports, doctors can provide feedback and arrange new rehabilitation plans for their patients, which are carried out in the application for doctors and sent to patients through the web server.

In summary, our main contributions in this paper include:

- We create the first lower body rehabilitation human key points detection dataset, which focuses on the recognition of lower body rehabilitation exercises, and helps to improve the neural network model performance for pose estimation on rehabilitation activities.
- We propose a lightweight human pose estimation model that runs on a smartphone smoothly without sacrificing much in accuracy.
- We design an in-home lower body rehabilitation system that allows patients to carry out rehabilitation by themselves at home through a smartphone. Evaluation reports for rehabilitation activities of a patient are sent to his/her primary doctors so that they can follow up on the patient's progress and arrange new rehabilitation activity plans.

II. Related work

A. Lower body rehabilitation

The past decade has seen the rapid development of lower body rehabilitation in many cases. After knee arthroplasty, ambulation recovery is the primary concern for lower body rehabilitation patients. This requires a series of knee exercises that enable patients to improve the range of motion easily and carry out activities of daily life [20].

Along with the growth in clinical rehabilitation, however, there are increasing concerns over the cost and efficiency. The main challenge faced by clinical rehabilitation is that

the recovery usually calls for a long-term intensive exercise program, which is time-consuming, expensive, and difficult. Motivated by this, several studies have been conducted on in-home rehabilitation. For instance, [21] tracks patients' movements via video capture virtual reality technology to reduce the cost while increasing efficiency. To further enhance patient engagement, [22] shows that well-designed video games such as motion-controlled video games could be an effective supplement to traditional physical therapy.

B. Human pose estimation

Human pose estimation remains one of the hottest research topic for decades. Before the advent of CNN, people have devised a variety of features to detect body key joint points in images [3] [4] [5]. After the AlexNet [6] won the ImageNet challenge in 2012, CNNs have been widely used in human pose estimation [7] [8] [9] [10] [11] [12] [13] [26]. For example, [26] proposes a Cascade Pyramid Networks to combine feature information from multiple scale representation maps. HRNet [13] keeps a high resolution feature representation branch through the whole architecture. The high resolution features will be augmented with lower resolution features. Benefiting from multiple resolution features, HRNet achieves the state-of-the-art performance in major public datasets, such as COCO keypoint detection dataset [14], MPII [15], and PoseTrack [16].

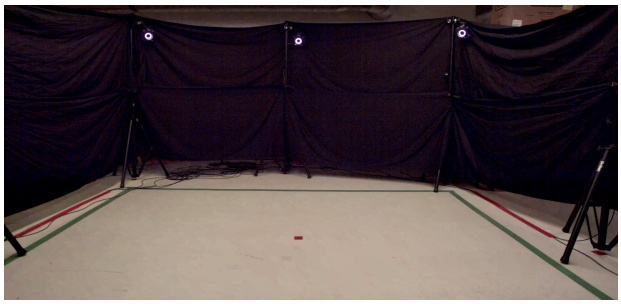
C. Human body key points detection dataset

The COCO Keypoint Detection Dataset [14] has gone through different versions since its creation. The latest popular version was published in 2017. It contains more than 57K, 5K and 20K images for training, validation, and test respectively. The MPII Human Pose dataset [15] contains 25K images with more than 40K subjects, in which there are 28K subjects for training, and the rest for testing. The extended LSP dataset [1] consists of 11K training images from sports activities and 1K images for testing. All images in these datasets are collected from real-world scenarios across a wide range of activities. However, these activities are very different from rehabilitation exercises. The Human3.6M dataset [17] provides more than 3.6M images with labeled key points generated indoor by a motion tracking system from several volunteers. However, this dataset only covers several special daily activities, which are also different from rehabilitation activities.

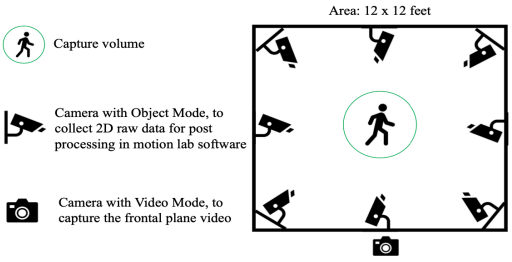
III. Our Method

A. System description

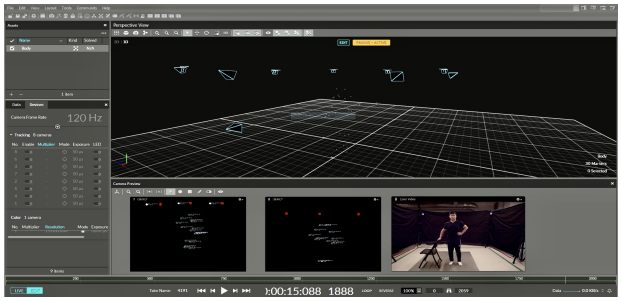
Our system is based on the latest deep learning algorithms tailored for human pose estimation. It can identify and track the movement of human joint points captured in live video or stored video files, without requiring any wearable accessories on the human body. Given the coordinates of the joints, we can calculate the angle of the target joints and evaluate the activities of the patients. As shown in Fig. 1, our research is mainly divided into three phases. In the first phase, we collect and preprocess data to construct Lower Body Rehabilitation



(a)



(b)



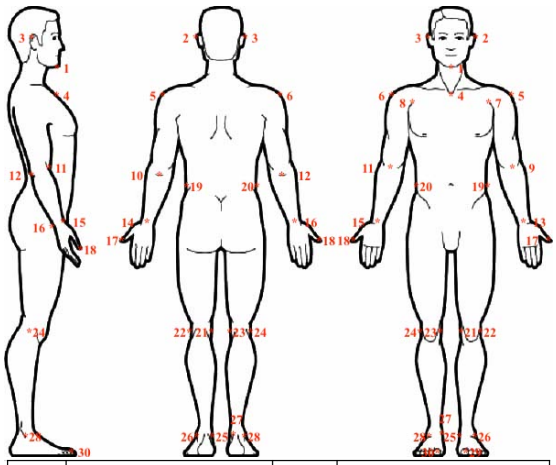
(c)

Fig. 2: (a) Motion Lab environment. (b) Cameras placement. (c) Motion tracking system screenshot.

Dataset. Following this, we train the proposed Lightweight Pose Estimation Model and then deploy it in the third phase. Our prototype system can be deployed on a mobile platform as a Health Insurance Portability and Accountability Act (HIPPA)-compliant app, allowing greatest flexibility in location and time of physical therapy rehabilitation exercises.

B. Lower body rehabilitation dataset

In the research for the rehabilitation assessment, it is necessary to build a specific dataset for lower body rehabilitation. First of all, we collect data in a motion lab, which consists of nine high-performance cameras and an optical motion tracking system (Fig. 2). Each camera covers a wide angle of 70 degrees view with a resolution of 1.7 MP, offering an expansive camera coverage with a capture rate of 360 FPS [2]. The optical motion capture system is equipped with the tracker's industry-leading 3D reconstruction and rigid body solution. As shown in Fig. 3, we attach reflective markers on the human body joints axes. The nine cameras are placed around the human body, eight of which continuously track the movement of markers in "Object Mode". In such a way, these eight cameras record raw 2D video frames, which are then processed by Motive software to generate 3D coordinates of each marker.



Index	Position of Markers	Index	Position of Markers
1	Chin	16	RWrist_dorsal
2	Ear lobe_Left	17	Thumb_Left
3	Ear lobe_Right	18	Thumb_Right
4	Chest-sternal notch	19	Waist_Left(waist belt)
5	Shoulder acromion_Left	20	Waist_Right(waist belt)
6	Shoulder acromion_Right	21	LKnee_medial condyle
7	Coracoid_Left	22	LKnee_lateral condyle
8	Coracoid_Right	23	RKnee_medial condyle
9	LElbow_cubital fossa	24	RKnee_lateral condyle
10	LElbow_olecranon	25	LAnkle_medial malleolus
11	RElbow_cubital fossa	26	LAnkle_lateral malleolus
12	RElbow_olecranon	27	RAnkle_medial malleolus
13	LWrist_ventral	28	RAnkle_lateral malleolus
14	LWrist_dorsal	29	Toe_Left big toe
15	RWrist_ventral	30	Toe_Right big toe

Fig. 3: Marker Position

The remaining camera is placed in front of the human body at a height of 60 inches, perpendicular to the plane of interest, and is responsible for recording the frontal RGB video of the target. As a result, we obtain the ground truth 2D/3D locations of all the markers through the motion tracking system.

The main challenge faced during the data collection is the shift of markers on the body due to movement of limbs. Proper marker placement is vital for the quality of motion data because each marker on a tracked subject is used as indicators for both position and orientation. For the purpose of mounting marker on skin, we adopt the rigid plastic marker base to reinforce the stability of marker. Another challenge is the motion tracking software does not label the tracked markers automatically. Therefore we have to manually label each detected marker with an accurate keypoint frame by frame in the videos. With the motion tracking system, we are able to label markers through several consecutive time frames.

C. Clinical requirements of lower body rehabilitation exercises

In our data collection process and later evaluation, we take into consideration the clinical requirements of lower body rehabilitation as advised by a licensed physical therapist. This is because (1) we aim to ensure that the exercise performed by volunteers meet the standard of rehabilitation purposes; and (2) the assessment on the rehabilitation exercise images relies on the clinical guidelines and common practice with respect to quantitative metrics such as range of motion and angle of limbs.

Activity Name	Activity Type
Supine Ankle Pumps	**Supine**
Small Range Straight Leg Raise	
Supine Alternating Small Range Straight Leg Raise	
Supine Short Arc Quad	
Supine Bridge	
Supine Hip Abduction	
Supine Knee to Chest with Leg Straight	
Supine Heel Slides	
Normal Range Straight Leg Raise	
VMO Straight Leg Raise	
Sidelying Hip Abduction	
Mini Squat with Counter Support	**Standing**
Standing March with Counter Support	
Standing Hip Abduction	
Standing Hip Adduction	
Standing Hip Flexion with Chair Support	
Standing Marching	
Standing Hip Extension	
Step Up	
Step Down	
Lateral Step Ups	
Seated Ankle Circles	**Seated**
Seated Ankle Pumps	
Seated Active Assistive Knee Extension and Flexion Foot on Floor	
Sit to Stand	
Squat with Chair Touch	
Seated Long Arc Quad	
Seated March	
Seated Hip Flexion	
Seated Knee Flexion Extension AROM	
Seated Heel Raise	

TABLE I: Rehabilitation Activities in our study. It contains three main categories: Supine(11 exercises), Standing(10 exercises), and Seated(10 exercises).

Under the guidance of a licensed physical therapist, we select the most common and widely adopted 31 therapeutic exercises (Table I) in our study. These exercises can be divided into 3 categories: Supine, Standing, and Seated. During data collection, we strictly follow the standards provided by therapists. Some details are described as follows.

For the speed of motion, all exercises are recommended to be performed slowly which is rule number one for a home exercise program, especially in the early stage of recovery. In a newly operated total knee replacement, the knee Range of Motion (ROM) is usually limited. This is particularly true for patients after the knee replacement operation. Take Seated Heel Raise as an example, patients only need to lift the heel off the ground to be considered as completed.

For the range of motion, note that in Table I, actions can be very similar to each other. For instance, Small Range Straight Leg Raise requires a 30 to 45 degrees raise of legs while Normal Range Straight Leg Raise normally asks for raising leg from 45 to 75 degrees. Meanwhile, for these two actions, it is important to point the toes up in the exercising leg. The inactive leg must be bent on the bed to protect the patient's lower back. Such quantitative metrics are counted during rehabilitation activity evaluation.

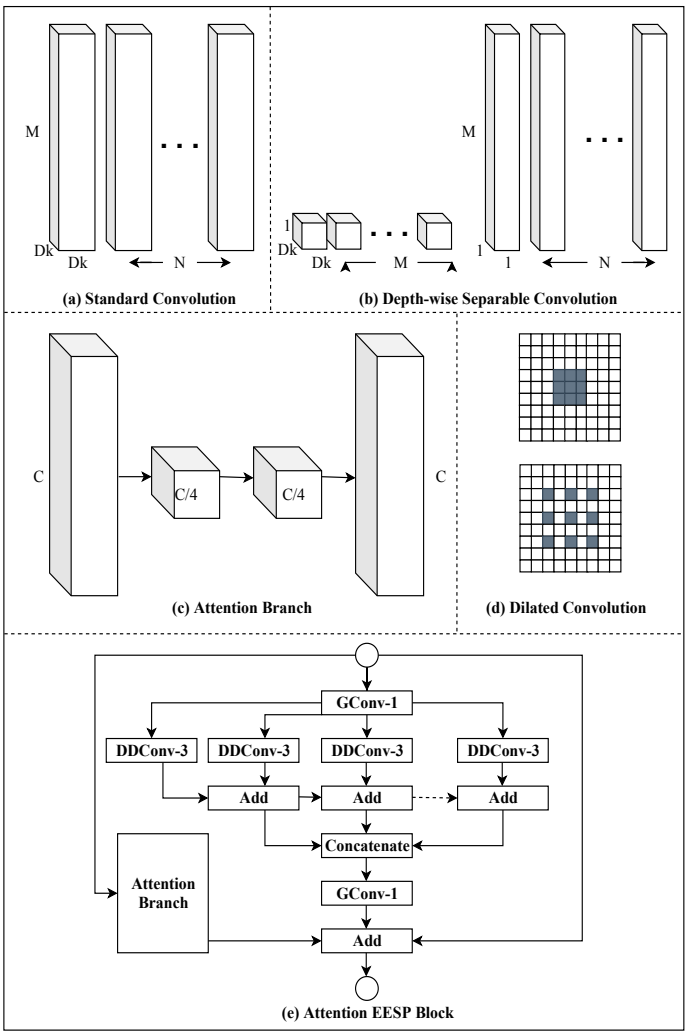


Fig. 4: Illustrating the details used in our proposed model.

For the activity type of standing and sitting, the upper body usually needs to be kept upright. This can be assisted by supporting objects. For example, in the activity of Sit to Stand and Seated Hip Flexion, patients can support their trunks by holding a chair with their hands to keep their bodies stable. In addition, another rule is that a patient should move his/her hip, knee and ankle joints in normal alignment. This rule applies to Supine Heel Slides, Sidelying Hip Abduction, Seated Active Assistive Knee Extension and Flexion Foot on Floor, Standing Hip Flexion with Chair Support, Seated Long Arc Quad, and Seated Knee Flexion Extension AROM, etc.

The range of motion angles and the body position are important indicators to a physical therapist to evaluate the compliance and effectiveness of rehabilitation exercises. We strive to pay attention to these metrics and requirements during the data collection process where volunteers are trained and instructed to perform in the motion lab. Our goal is to use such a dataset to train neural network models which can be utilized to recognize the keypoints and calculate range of motion and assess compliance to body position requirements.

In the data collection process, each volunteer performs

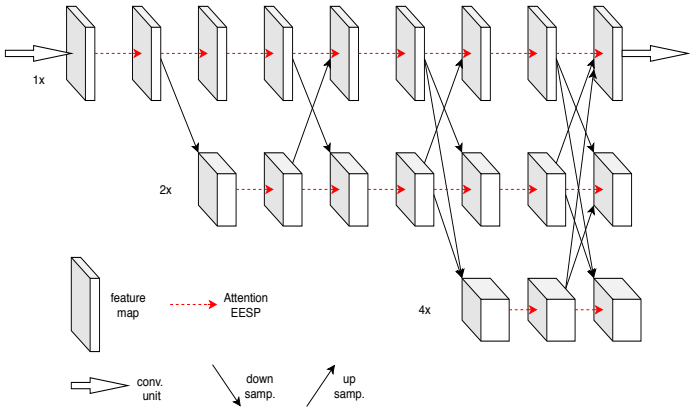


Fig. 5: Illustrating the architecture of the proposed model. It keeps the multiple resolution subnetworks, down samplings and up samplings from the original HRNet. But the deep convolutional blocks are replaced by the attention EESP blocks in Fig. 4(e). Better viewed in color.

31 activities that lead to about 53 videos recorded (some require both left and right camera views), and the duration of each video is approximately 25 seconds. Generally, for each volunteer, it takes about 2 hours to collect satisfactory movement data as the volunteers need to be trained. After collecting raw 2D/3D data of each person, it takes about 10 hours to complete the post-processing which requires extensive time to label the keypoints in video frames based on marker locations.

D. Lightweight human pose estimation model

We make effort to design a neural network model that is both accurate and lightweight in computation. MobileNet [18] is the most widely adopted deep convolutional neural network backbone for computer vision applications deployed on mobile devices. In [28], the authors implement a real-time human pose estimation model running on CPU based on MobileNetv1 [18], however, with moderate performance. Meanwhile, other deep CNN backbones for pose estimation, such as Hourglass [8], ResNet [12] and HRNet [13], have achieved great performance. In particular, HRNet leverages multiple resolution branches, which keeps the high resolution branch over the whole network, and gradually adds a lower resolution branch in each following stage. At the end of each stage, the feature maps from all other lower resolution branches are merged to the high resolution branch. Through this procedure, the HRNet has excellent representation capability. Nevertheless, HRNet network is too complicated to be deployed on mobile devices.

To benefit from the advantage of the HRNet and the MobileNet, we adopt HRNet as the backbone and leverage the modules in MobileNetv1 to simplify the original network. Specifically, the depth-wise separable convolution was first proposed in MobileNet [18], which is one of the most common choices in lightweight deep CNN models recently. The process of the standard convolution is showed in Fig. 4(a). While the two steps of the depth-wise separable convolution are showed

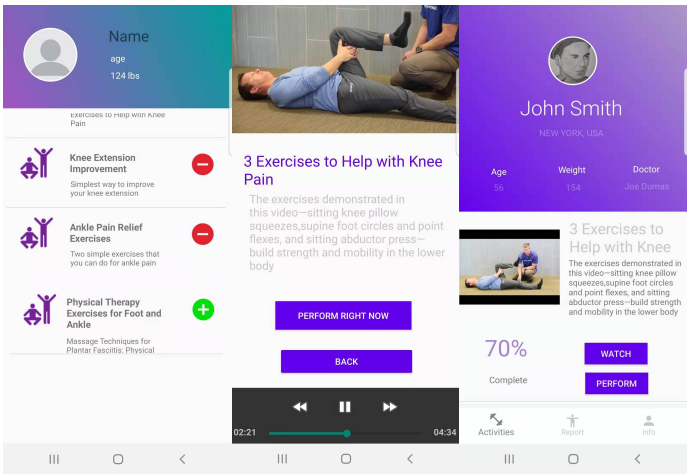


Fig. 6: MotionPal User Interfaces (left to right: Assign Exercise Action, Preview Exercise Video, Perform Exercise Action)

in Fig. 4(b). The computation cost for the standard convolution in Fig. 4(a) is calculated by (assuming padding is used):

$$D_K \times D_K \times M \times N \times D_F \times D_F \tag{1}$$

where $D_K \times D_K$ is the kernel size, $D_F \times D_F$ is the feature map size, M is the number of input channels and N is the number of output channels. In contrast, the computation cost for the depth-wise separable convolution in Fig. 4(b) is:

$$D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F \tag{2}$$

Comparing these two, we get a reduction in computation of:

$$\frac{D_K \times D_K \times M \times D_F \times D_F + M \times N \times D_F \times D_F}{D_K \times D_K \times M \times N \times D_F \times D_F} = \frac{1}{N} + \frac{1}{D_K^2} \tag{3}$$

In this paper, we use 3x3 filters in depth-wise separable convolution which saves about 90% of the computation compared to the standard convolution.

The attention mechanism has demonstrated salient advantages in the computer vision community [24]. In this paper, we design a special channel attention module as illustrated in Fig. 4(c). In this module, the feature map's channels are compressed to a quarter of the original size before they are recovered to the original number of channels, which generates attention-aware features. All these operations are implemented by 1x1 convolutional layers.

We also leverage the dilated convolution [23] for its spatial features extraction capability, which is a technical improvement over the standard convolution to extract feature information from a wider area of the input with the same computation cost. In Fig. 4(d), the bottom block shows the process of the dilated convolution with a rate of 2 compared with the standard convolution at the top.

Combining the modules above, we obtain a novel Attention EESP (Extremely Efficient Spatial Pyramid of Depth-wise Dilated Separable Convolutions [19]) block that not only uses much fewer parameters and computation operations than those in the original HRNet architecture but also processes excellent representation capability. The original EESP unit splits the standard convolution to several branches and chooses depth-wise separable convolutions to reduce most of the computational cost. Meanwhile, it adopts the dilated convolutions to improve the feature representation capability. We redesign the EESP unit by adding a channel attention branch. Fig. 4(e) shows the architecture of our attention EESP block. As shown in Fig. 4(c), the convolutional filters used in our attention branch are of 1x1 size, which has a computational cost of:

$$2 \times C \times \frac{C}{4} \times D_F \times D_F + \frac{C}{4} \times \frac{C}{4} \times D_F \times D_F = \frac{9}{16} \times C \times C \times D_F \times D_F \tag{4}$$

where $D_F \times D_F$ is the feature map size and C is the input channel. The cost is much smaller than that of the standard convolution in Equation (1).

Fig. 5 shows the architecture of our model. The red arrows represent our attention EESP blocks as shown in Fig. 4(e), and the rest are the same as in HRNet.

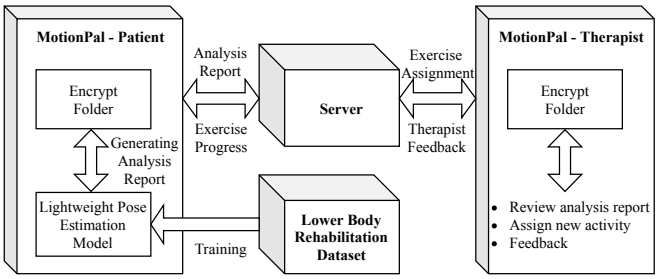


Fig. 7: MotionPal Application Architecture

E. MotionPal Application

Leveraging the refined neural network model, we have designed an application MotionPal on a mobile platform to perform the rehabilitation motion capture and analysis functionalities without any wearable accessories (Fig. 7). It executes our novel deep learning algorithm and provides insights into the design of a new scenario of in-home rehabilitation. More importantly, because we introduced different roles and interfaces for therapists and patients, MotionPal provides a bridge between therapists and patients for evaluating the rehabilitation outcomes.

MotionPal allows a patient to perform rehab exercise at home following the instructions and receive an analysis report based on the trained model. The only preparatory work a patient has to do is placing the smartphone at the optimal recording location at a height of 60 inches, as same as the position of our video mode camera during the data collection process. The analysis report will be shared with the therapist responsible for the patient. As shown in Fig. 6, the screenshots present the main features of MotionPal. A patient client can use the app to preview the exercise activity video, perform the exercise activity, review the analysis detail report, and modify

Method	Backbone	Pretrain	Input size	#Params	GFLOPs	AP	AP^{50}	AP^{75}	AP^{M}	AP^{L}	AR
8-stage Hourglass [8]	8-stage Hourglass	N	256x192	25.1M	14.3	66.9	-	-	-	-	-
CPN [26]	ResNet-50	Y	256x192	27.0M	6.20	68.6	-	-	-	-	-
SimpleBaseline [12]	ResNet-50	Y	256x192	34.0M	8.90	70.4	88.6	78.3	67.1	77.2	76.3
SimpleBaseline [12]	ResNet-101	Y	256x192	53.0M	12.4	71.4	89.3	79.3	68.1	78.1	77.1
SimpleBaseline [12]	ResNet-152	Y	256x192	68.6M	15.7	72.0	89.3	79.8	68.7	78.9	77.8
HRNet-W32 [13]	HRNet-W32	N	256x192	28.5M	7.10	73.4	89.5	80.7	70.2	80.1	78.9
HRNet-W32 [13]	HRNet-W32	Y	256x192	28.5M	7.10	74.4	90.5	81.9	70.8	81.0	79.8
HRNet-W48 [13]	HRNet-W48	Y	256x192	63.6M	14.6	75.1	90.6	82.2	71.5	81.8	80.4
SimpleBaseline [12]	ResNet-152	Y	384x288	68.6M	35.6	74.3	89.6	81.1	70.5	79.7	79.7
HRNet-W48 [13]	HRNet-W48	Y	384x288	63.6M	32.9	**76.3**	90.8	**82.9**	**72.3**	**83.4**	**81.2**
Lightweight OpenPose [28]	MobileNetv1	N	384x288	**2.5M**	3.6	62.5	87.3	67.8	58.6	68.0	72.3
LPN [29]	ResNet-50	N	256x192	2.9M	**1.0**	69.1	88.1	76.6	65.9	75.7	74.9
Ours	HRNet-W32	N	256x192	4.7M	**1.0**	70.8	**91.5**	78.3	68.2	74.6	73.8

TABLE II: Comparisons on the COCO validation set. Pretrain = pretrain the backbone on the ImageNet classification task.

Method	Backbone	#Params	GFLOPs	AP
SimpleBaseline [12]	ResNet-50	34.0M	8.90	87.0
HRNet-W32 [13]	HRNet-W32	28.5M	7.10	**89.3**
Ours	HRNet-W32	**4.7M**	**1.0**	88.3

TABLE III: Comparisons on our Lower Body Rehabilitation Activities Keypoint Detection validation set. The input size is 256x192 for all models.

Method	#Params	GFLOPs	AP
HRNet-W32 [13]	28.5M	7.10	73.4
HRNet-W32 + EESP block	3.2M	0.65	65.9
HRNet-W32 + Attention EESP block s=2	7.8M	1.8	71.2
HRNet-W32 + Attention EESP block s=4	4.7M	1.0	70.8
HRNet-W32 + Attention EESP block s=8	3.8M	0.79	67.9

TABLE IV: Comparisons on the COCO validation set. The input size is 256x192 for all models.

the basic health information. For a therapist, MotionPal allows him/her to review the patients list, the basic health information, the detailed analysis report, and the activities progress made by each patient. Moreover, the therapist is able to assign new exercise activity to patients.

This application was built to be compliant with the Health Insurance Portability and Accountability Act (HIPAA), including strong password protection and 128-bit encryption. In fact, as the video capture and analysis are executed locally on the smartphone and all of the data will be stored in the smartphone of the user, there is no protected health information (PHI) that is transferred outside from the device.

IV. Experiments

We evaluate our proposed model on both public dataset (i.e. COCO2017) and our new lower body rehabilitation dataset. In this way, we can understand the generality and effectiveness of the model.

A. COCO Keypoint Detection

Dataset. COCO keypoint detection dataset [14] contains over 200K images and 250K person instances labeled with 17 keypoints. We train our model on COCO train2017 dataset that consists of 57K images and 150K person instances. We evaluate our model on the val2017 set with 5000 images.

Evaluation metric. The keypoint evaluation metrics used by COCO is Object Keypoint Similarity (OKS): OKS = $\Sigma_i[\exp(-d_i^2/2s^2k_i^2)\delta(v_i > 0)]/\Sigma_i[\delta(v_i > 0)]$. Here d_i is the Euclidean distance between each corresponding ground truth and detected keypoint, v_i is the visibility flag of the ground truth, s is the object scale, and k_i is a per-keypoint constant that controls falloff. We report standard average precision and recall scores [25]: AP (AP at OKS=.50:.05:.95, primary challenge metric), AP^{50} (AP at OKS=.50, loose metric), AP^{75} (AP at OKS=.75, strict metric); AP^{M} (AP for medium objects) , AP^{L} (AP for large objects); and AR at OKS=.50:.05:.95.

Training and Testing. We follow most of the default settings in HRNet [13] including the input size, data augmentation, and the Adam optimizer. But the learning schedule is different, which sets the initial base rate as 1e-2, and reduces it to 1e-3, 1e-4, and 1e-5 at the 20th, 170th, and 200th epochs, respectively. This schedule is chosen because we start the training from scratch. The training process is terminated within 210 epochs.

To test our model on the validation set, we adopt the two-stage top-down paradigm: detecting the person instance using a person detector, and then predicting keypoints. We use the same person detectors provided by HRNet [13] for the validation set.

Results on the validation set. We compare the results of our model and other state-of-the-art methods in Table II. Our model adopts HRNet-W32 as the backbone. We train the model from scratch with an input size of 256x192, which achieves an AP score of 70.8, with 4.7M parameters and only 1.0G FLOPs. This result outperforms other lightweight models, such as Lightweight OpenPose [28] and LPN [29]. Especially, our model achieves the best AP^{50} score.

B. Lower Body Rehabilitation Activities Keypoint Detection

Dataset and evaluation metric.. Our lower body rehabilitation activities keypoint detection dataset is organized following the COCO dataset. Our dataset is split into a training set and a validation set of 500,000 images and 10,000 images, respectively. And the images in the training set and the validation set are from different volunteers. The ground truth results are stored in JSON files in the same structure as in the COCO dataset. There are 30 keypoints for each person

instance in our dataset, and there is only one person in each image. All the images are of 720p resolution with the same size. The evaluation metric used in our dataset is the same as in the COCO dataset.

Training and Testing. Firstly, we modify the COCO API codes to process the images in our dataset and calculate the standard average precision and recall scores. Then we follow the same input size, data augmentation, Adam optimizer, and the learning schedule as for the COCO dataset. The input size is 256x192 for all models in our experiments. And all models are trained from scratch.

Results on the validation set. We compare the results of our model and the benchmark methods in Table III. While the original HRNet-W32 [13] network achieves the best performance with an 89.3 AP score, our model performs slightly worse with an AP score of 88.3, better than the SimpleBaseline [12] with ResNet-50. Note that our model size (#Params) is 16% of HRNet-W32 and the complexity (FLOPs) is 14%.

C. Ablation Study

To investigate the effectiveness of our architecture, we carry out the ablative analysis on the COCO validation set so that the results can be compared with prior work. We compare the models with and without different types of attention branches. In our attention branch, we first compress the number of channels to c/s, where c is the number of channels in the input feature map, s is the scale parameter. After a convolutional layer, the number of channels is recovered to c. Firstly, the original EESP blocks without attention branches are used to replace the deep convolutional blocks in the HRNet-W32. Then, we test our attention EESP block with different scales for channel attention branches. All these models are trained from scratch on COCO training set with the input size 256x192 and tested on COCO validation set. As shown in Table IV, all the models using attention branches achieve better AP scores than that without attention branches.

V. Conclusion

In this paper, we propose a human pose estimation based lower body rehabilitation system to assistant patients with lower body rehabilitation activities at home by themselves only through a smartphone. We also redesign a lightweight deep CNN model for human pose estimation that runs on a mobile device smoothly. Experiment results show that our model, using much fewer parameters and less computation cost, achieve comparable performance with the state-of-the-art method.

The future work includes improving the deep CNN model for mobile devices, extending the rehabilitation detection dataset to contain more types of rehabilitation activities from more patients, and increasing the 3D location information of the human body keypoints for 3D pose estimation.

References

[1] S. Johnson, M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation," In British Machine Vision Conference, volume 2, page 5, 2010.

[2] OptiTrack. (2020, Jan.) "Prime 17W In Depth".[Online]. Available: https://optitrack.com/products/prime-17w/indepth.html

[3] V. Ferrari, M. Marin-Jimenez, A. Zisserman, "Progressive search space reduction for human pose estimation," CVPR, 2008.

[4] M. Andriluka, S. Roth, B. Schiele., "Pictorial Structures Revisited: People Detection and Articulated Pose Estimation," CVPR, 2009.

[5] Preben Fihl, Thomas B. Moeslund, "Pose Estimation of Interacting People using Pictorial Structures," Advanced Video and Signal Based Surveillance (AVSS) 2010 Seventh IEEE International Conference on, pp. 462-468, 2010.

[6] Krizhevsky, A., Sutskever, I., Hinton, G.E., "Imagenet classification with deep convolutional neural networks," Advances in Neural Information Processing Systems(NeurIPS), pp. 1097–1105, 2012.

[7] Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y., "Convolutional pose machines," CVPR, 2016.

[8] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," ECCV, 2016.

[9] Cao, Z., Simon, T., Wei, S.E., Sheikh, Y., "Realtime multi-person 2d pose estimation using part affinity fields," CVPR, 2017.

[10] Yang, W., Li, S., Ouyang, W., Li, H., Wang, X., "Learning feature pyramids for human pose estimation," IEEE International Conference on Computer Vision(ICCV), 2017.

[11] Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J., "Cascaded pyramid network for multi-person pose estimation," CVPR, 2018.

[12] B. Xiao, H. Wu, and Y. Wei, "Simple baselines for human pose estimation and tracking," ECCV, pages 472–487, 2018

[13] K. Sun, B. Xiao, D. Liu, J. Wang, "Deep High-Resolution Representation Learning for Human Pose Estimation," arXiv:1902.09212, 2019

[14] T. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: common objects in context," ECCV, pages 740–755, 2014.

[15] M. Andriluka, L. Pishchulin, P. V. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," CVPR, pages 3686–3693, 2014.

[16] U. Iqbal, A. Milan, and J. Gall, "Posetrack: Joint multi-person pose estimation and tracking," CVPR, pages 4654–4663, 2017.

[17] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments," PAMI'13, 2013.

[18] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," arXiv:1704.04861, 2017.

[19] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi, "Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network," CoRR, abs/1811.11431, 2018.

[20] Shakespeare D, Kinzel V., "Rehabilitation after total knee replacement: time to go home?" Knee. 2005;12(3):185-189.

[21] Weiss PL, Rand D, Katz N, et al. "Video capture virtual reality as a flexible and effective rehabilitation tool," Journal of NeuroEngineering and Rehabilitation. 2004;1(1):12.

[22] Lohse KR, Shirzad N, Verster A, Hodges NJ, Van der Loos HFM, "Videogames and rehabilitation: Using design principles to enhance patient engagement," Journal of Neurologic Physical Therapy. 2013. 37: 166–175.

[23] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," ICLR, 2016.

[24] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, "Residual attention network for image classification," CVPR, 2017.

[25] MSCOCO. (2020, Jan.) "COCO Keypoints Evaluation". [Online]. Available: http://cocodataset.org/#keypoints-eval.

[26] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation", CVPR, 2018.

[27] ColoringSky. (2020, Jan.) "Human Body Front And Back Coloring Pages". [Online]. Available: https://www.coloringsky.com/human-body-front-and-back-coloring-pages.

[28] D. Osokin, "Real-time 2D Multi-Person Pose Estimation on CPU: Lightweight OpenPose", arXiv:1811.12004, 2018.

[29] Z. Zhang, J. Tang, G. Wu, "Simple and Lightweight Human Pose Estimation", arXiv:1911.10346, 2019.