Informatics in Medicine Unlocked 23 (2021) 100544

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: http://www.elsevier.com/locate/imu

A framework for automatic hand range of motion evaluation of rheumatoid arthritis patients

Luciano Walenty Xavier Cejnog [a,*], Teofilo de Campos [b], Valéria Meirelles Carril Elui [c], Roberto Marcondes Cesar Jr. [a]

[a] Departamento de Ciência da Computação, IME-USP, Rua do Matão, 1010, São Paulo, Brazil
[b] Departamento de Ciência da Computação, Universidade de Brasília, Brazil
[c] Departamento de Terapia Ocupacional, FMUSP Ribeirão Preto, Brazil

ARTICLE INFO

Keywords:
Hand pose estimation
Computer vision
Hand occupational therapy
Depth images

ABSTRACT

We propose a framework for evaluation of finger movement patterns on Rheumatoid Arthritis patients: flexion, extension, abduction and adduction. The framework uses a state-of-the-art 3D hand pose estimation method that runs in real-time, allowing users to visualize 3D skeleton tracking results at the same time as the depth images are acquired. We compute flexion and abduction angles from the obtained skeleton pose parameters. We performed data acquisition from a cohort of patients and a control set and compared the angles from those two sets of people. An analysis using time series similarity with frequency domain descriptors is adopted to characterize the movement patterns for flexion/extension. We performed classification experiments using these descriptors, thus distinguishing movement sequences of hands with rheumatoid arthritis from healthy hands. The descriptors used in the classification experiment were effective and reached average results of 89% in scenarios of unseen subjects, and an average of 82% in experiments with sample synthesis that allow a more robust statistical performance evaluation. Our framework allows the characterization of the current state of the disorder in each patient, with minimal intervention and reduced evaluation time.

1. Introduction

Contactless technology is an important trend in biomedical and health informatics [1]. Among the different advances, it is worth mentioning smart personalized healthcare and telemonitoring [2]. The application of AI and machine learning is of particular interest in order to fight pathologies such as neurodegenerative disorders and degenerative arthritis, for instance Ref. [3]. This paper introduces a computer vision approach to analyze movement patterns from patients on hand occupational therapy. We focus on rheumatoid arthritis (RA) recovery.

RA is a chronic autoimmune disease that leads to joint deformities due to an inflammation that causes the erosion of tissues, including bones. This inflammatory mechanism was discovered very recently [4]. Findings of population-based studies show RA affects 5–10% of adults in developed countries. The disease is three times more frequent in women than men, and 50% of risk of developing RA is attributable to genetic factors [5]. The clinical complaints include pain, swelling and motion limitations of the affected joints. A physical examination reveals the presence of pain, increased joint volume, intra-articular effusion, heat and eventual redness [6]. Its complications can lead to deformity and destruction of joints, due to the erosion of bone and cartilage. Although the risk of death is nonexistent, RA severely affects the quality of daily life of the patients. In the hand, deformities can reach all articulations, causing subluxations and deformities in metacarpophalangeal joints, interphalangeal joints and wrists, affecting motor functions. In some cases, the progression of the deformity causes ulnar deviation, with the destruction of the wrist ligaments that move towards radial deviation. Using body compensatory mechanisms, the excessive forces are transferred to the fingers' extensor tendons [7]. Fig. 1 shows an example of a hand with ulnar deviation, in contrast with a healthy hand.

Since RA is a chronic disease, early diagnosis is important for preventing the progression of the deformities, and it is important that persistent joint inflammation, progressive joint damage and continuing functional decline are assessed by the therapists. Typically, *Disabilities of the Arm, Shoulder and Hand* (*DASH*) questionnaires are used to assess hand function during the recovery process. This evaluation method is

* Corresponding author.
E-mail addresses: cejnog@ime.usp.br (L.W.X. Cejnog), rmcesar@ime.usp.br (R.M. Cesar Jr.).

https://doi.org/10.1016/j.imu.2021.100544
Received 17 September 2020; Received in revised form 24 February 2021; Accepted 24 February 2021
Available online 3 March 2021

Fig. 1. Example of hand with finger ulnar deviation (on the right) in contrast with a normal hand (on the left) of the same person. Such hand shapes represent a challenge for hand tracking algorithms.

based on the patient qualitative self-evaluation of difficulty in the execution of daily activities, such as to write, to prepare a meal or to make a bed. Quantitative evaluation uses range of motion measurements. The standard procedure for measuring range of motion angles involves teh use of a goniometer. This procedure is easy and relatively reliable, however it is time-consuming and in cases with more deformities can be difficult and tiring to the patient. In this paper, we propose a camera-based framework that can enhance the comfort of the patient and the efficiency during the range of motion angle assessments. Our procedure is markerless, does not require environmental setups and uses state-of-art computer vision techniques. Although this is a very important health problem, there are few computer vision methods described in the literature to automatically analyze the treatment evolution and assess the patient range of motion. This paper presents steps to fill this gap.

The proposed methodology follows the pipeline shown in Fig. 2. Our system handles both left and right hands with ulnar deviation, as well as healthy hands. The pipeline starts with data acquisition using RGBD sensors from a number of patients. A key step in the proposed approach is to accurately locate hand joints in 3D. For each depth image, a state-of-art 3D hand pose estimation method is applied, producing a 3D skeletal model of 21 joints. These skeleton poses (which are represented as an array of 3D points) are then analyzed in order to estimate flexion and abduction angles and range-of-motion measurements, that should be used by the therapist in the treatment. We further analyze the signals obtained with sequences of angles to identify structural patterns. We classify flexion movements into patients and control set in order to show that the use of a state-of-art hand pose estimation algorithm is able to generalize for any hand pose given the ideal acquisition conditions, and thus such analysis is feasible.

The main contributions of this study are: (1) we introduce a new computer vision framework to support hand occupational therapy based on state-of the-art hand pose estimation; (2) we introduce hand movement analysis tools based on the estimated angles and range-of-motion measurements from skeletons; and (3) we present a new dataset of depth maps and hand tracking results obtained using from patients of Rheumatoid Arthritis being treated in one of the hospitals of the University of São Paulo[1].

This paper is a follow up from the analysis published in Ref. [8]. Here we present a new analysis based on time series similarity, with new classification results, including an evaluation of two feature extraction methods and several classifiers. The rest of this paper is organized as follows: Section 2 shows related works on hand pose estimation and range of motion measurements in occupational therapy; Section 3 details the proposed framework in each of its steps; Section 4 shows experimental results with the data and Section 5 presents concluding remarks.

2. Related works

2.1. Range of motion measurements

The evaluation of hand function is fundamental for the therapist to plan the treatment as well as to record the results. The literature in hand therapy defines metrics and guidelines in order to extract those measures with precision [9]. For measuring joint angles, a widely used metric is the range of motion (ROM), which consists in a set of angles between joints, whose maximum and minimum values are evaluated during flexion/extension and abduction/adduction movements. This is usually done in a static way or through mechanical sensors, hindering the ability to evaluate those angles while the patient is performing an occupational task. One of the most widely used assessment methods for range-of-motion measurement is goniometry. With a specific hand/-finger goniometer, the therapist can access objectively and reliably the range of motion measurements. The goniometer is widely used due to its simplicity and low cost. However the measurement procedure requires a trained therapist that must follow protocols. The task is manual, time consuming, repetitive, intrusive for the patient and prone to error. Also, the measurements are taken in a controlled scenario, and may not reflect the real abilities on execution of daily activities. DASH questionnaires are used to complement this evaluation. Despite these problems, comparing to traditional 2D visual estimation and wire tracing, goniometry shows more reliability and precision [10,11].

An alternative to goniometry is assessing the range of motion measurements from digital photogrammetry. This approach was mostly used by surgeons, and some recent works show that the reliability of this method has increased over the years [12]. However, the viability study presented by Meals et al. [13] shows that so far the use of digital photogrammetry has limited effectiveness for measuring hand joint angles in comparison to the manual goniometry. One of the main limitations of photogrammetry techniques is that the result is not immediately assessed: joints must be photographed and then measured. The work indicates future possibilities of using 3D scanning and video capture technology to the development of an automatic goniometer for the hand. Since the results of the hand pose estimation are available in real-time, we believe that this paper presents an important first step in this direction. Other alternatives in the evaluation are the use of electronic goniometers, like the torque-based Multielgon system [14].

Among recent works that propose solutions based on computer vision, Pereira et al. [15] proposes a smartphone accelerometer-based app to measure active and passive knee ROM in a clinical setting. However, it is much more challenging to apply such technique to measure hand joints ROM (for starters, one cannot use mobile phones for that). Methods based on 2D images for hand pose estimation are still not viable. An alternative is the use of depth sensors, and despite their recent rise in popularity, few works to date make use of such devices for this task. We highlight the work of Lima et al. [16], which is a system that uses information obtained by a Leap Motion sensor to estimate hand angles.

We expect that the significant advances in computer vision and hand pose estimation can lead to a series of advances in this specific field of application. The possibility of acquiring 3D frames and skeletons reduces most ambiguities found in 2D visual estimation, and its use in the treatment of patients can be far less intrusive than using goniometers.

[1] An anonymized version of the dataset has been made available as part of this work at http://vision.ime.usp.br/~cejnog/handanalysis/.

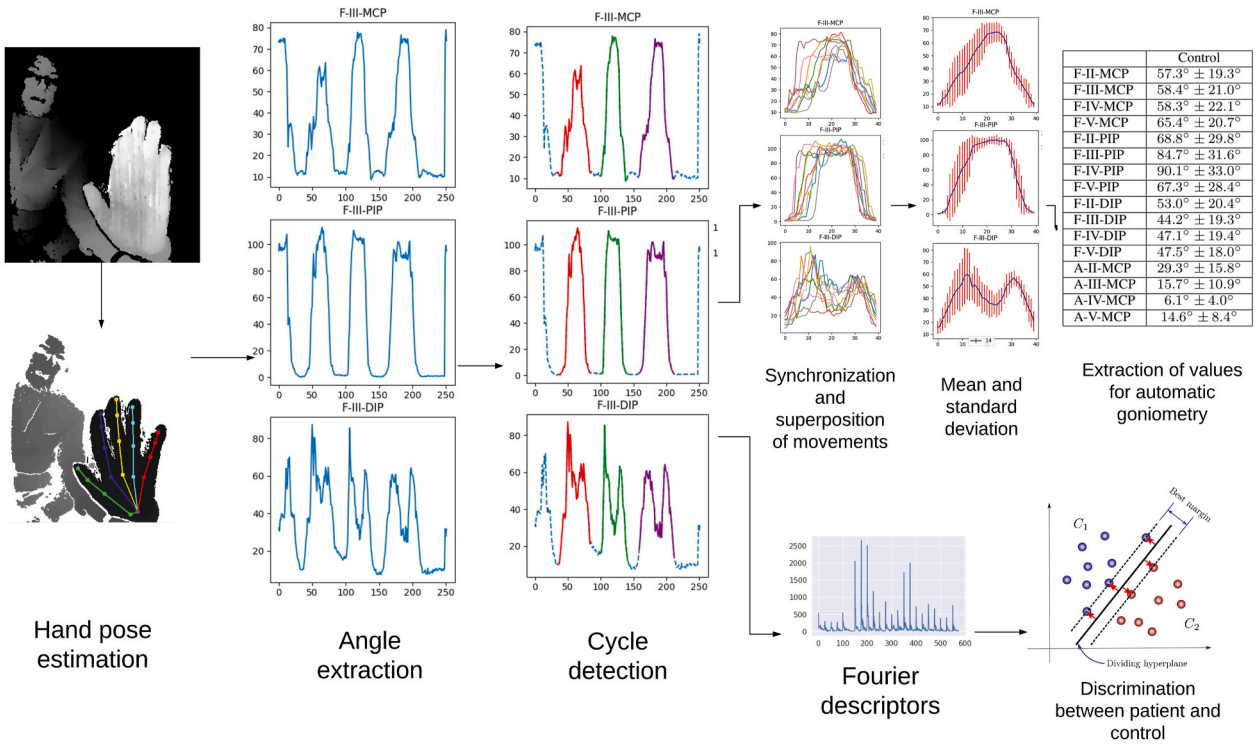

Fig. 2. The proposed pipeline extract and analyze hand joint angle measurements from RGBD images acquired in real-time.

2.2. Hand pose estimation

With the development of low-cost depth sensors, the most common input data for hand pose estimation is depth maps. The use of such sensors reduces the data ambiguity without the need of configuring and calibrating a multiple view setup, facilitating the hand pose estimation task.

In recent years, the development of deep learning algorithms led to significant advances in machine learning and its applications, particularly in Computer Vision. The advent of those algorithms combined with the development of accurate solutions for 2D joint detection based on Convolutional Neural Networks (CNNs) [17,18] led the community of 3D hand pose estimation to design methods that are also based on CNNs [19–24]. Those methods differ among themselves in the neural network architecture and type, the input image type, the hand representations used and the use of prior constraints. As an example, the *DeepPrior++* [20] uses a Residual Neural Network whose training is based on minimizing residual weights in each layer. It also uses data augmentation in which realistic samples were generated from simple geometric transformations on the original training samples. Guo et al. [24] use an ensemble-based neural network which integrates the results of different regressors in different regions of the image. Chen et al. [25] compute a feature map for each joint and fuse those maps using a structured region ensemble network (named Pose-REN), reaching solid results. Wan et al. [26] propose the combination of a Generative Adversarial Network (GAN) to model the distributions of depth maps and a Variational Autoencoder (VAE) to model the distribution of hand poses. This method allows training and learning from unlabeled data.

Fang et al. [27] recently proposed JGR-P2O, a system for pixel-to-offset predictions based on joint graph reasoning. This system explicitly models the dependencies among joints and the relations between pixels and the joints for better local feature representation learning. This method unifies pixel-wise offset predictions and direct joint regression for end-to-end training, leading to state-of-the-art results with a relatively low computational cost.

Another line of work includes methods based on volumetric information, which use context features of the 3D point sets in order to more accurately locate the joints. Methods such as A2J [28], V2V [29] and DenseNet [30] currently reach the best results in all state-of-art datasets for hand pose estimation. DenseNet obtains the hand pose by fusing 2D and 3D heatmaps. V2V uses an encoder-decoder architecture to convert the 2D depth image in a 3D voxel grid, and then estimates the per-voxel likelihood of each keypoint, identifying the positions of highest likelihoods. These are then warped back to real world coordinates. This approach has the drawback of the high computational cost of the voxelization procedure, increasing the difficulty of the training process. A2J uses anchor points in the depth image which capture the global-local context information. The joint position is regressed by weighting the influence of each anchor point. The neural network used is a 2D-CNN, which lowers the computational cost of training. This method currently reaches the best performance results on HANDS17 dataset [31], while JGR-P2O is the best performing method on NYU and ICVL datasets.

The current panorama of the area indicates that there is still room for improvement on methods based on deep CNNs using depth images, as this is the focus of many research groups around the world. This is particularly important in the application addressed in the present paper, since state-of-the-art methods trained on standard healthy hands tend to fail when applied to hands with deformities such as the one in Fig. 1.

Although a lot of improvement has been observed since the first commodity depth sensors became available, all methods have their limitations and depth maps are still far from perfect. One potential direction for future work is to exploit a pre-processing step to denoise depth maps using a method such as that of Yan et al. [32].

Another potential avenue to be explored for 3D hand pose estimation is the use of a tracking-as-detection approach (or indeed, tracking-as-retrieval), similar to what was done by Stenger and others in the mid-2000's [33]. Despite the relative success back then, such an approach does not seem to have been explored again since the deep learning revolution. The deep multi-view retrieval method of Yan et al. [34] has certainly a high potential of success in a tracking-as-retrieval framework.

Other problems that relate to inferring 3D from 2D RGB images, such as room layout estimation [35] can certainly inspire methods for hand pose estimation without depth information. More closely related are the methods of 2D joint detection based on CNNs and their successful application to problems such as human pose estimation [36]. Methods that use learning-based 2D joint detection and Inverse Kinematics have been proposed to estimate hand pose based exclusively on RGB image [37–40]. The development of monocular image-based pose estimation methods is important for generalization and ease of use, but the absence of the depth dimension makes the problem much harder. Data-driven methods need much larger datasets to be trained in order to obtain a good generalization ability.

3. Proposed approach

We follow the pipeline of Fig. 2, divided into data acquisition, 3D hand pose estimation and analysis. Table 1 presents a list of the main mathematical symbols used in this paper.

3.1. Data acquisition

As first step of the project, our goal was to acquire data from patients with hand deformities due to rheumatoid arthritis (RA). This data was obtained from patients of the Hospital das Clínicas from University of São Paulo. Several depth sensors were evaluated in a range of preliminary experiments and we found that the *Intel RealSense® SR300* generates the best depth maps and has a range which is the most suitable for our acquisition scenario.

The setup is illustrated in Fig. 3. For the acquisition, the distance between the camera and the elbow support was fixed, but the hand itself was kept free. However, the hand needs to be the nearest object with respect to the camera, otherwise the tracker produces very noisy results.

Table 1
Symbols and abbreviations used in the rest of the paper.

Symbol	Meaning
$\mathscr{D}$	Depth image
$\vec{S}(t)$	Skeleton at frame t
MCP_k	Metacarpophalangeal joint for finger k
PIP_k	Proximal interphalangeal joint for finger k
DIP_k	Distal interphalangeal joint for finger k
TIP_k	Tip joint for finger k
W	Wrist joint
CMC	Carpometacarpal thumb joint
$\widehat{\text{FMCP}}_k$	Flexion angle for joint MCP_k
$\widehat{\text{FPIP}}_k$	Flexion angle for joint PIP_k
$\widehat{\text{FDIP}}_k$	Flexion angle for joint DIP_k
ATIP_k	Abduction distance between tips of fingers k and $k-1$
$\vec{A} = a_i(t)$	Angle representation of a clip as a set of functions: a_i represents the ith. angle on frame t.
$\mathscr{F}_{a_i}$	Fourier transform of the angle representation a_i.
$\mathscr{F}$	Concatenation of Fourier transforms of all angles a_i that compose the angle representation $\vec{A}$ of the clip.
$\mathscr{N}(\mu,\sigma)$	Gaussian distribution with mean μ and standard deviation σ.
LOO	Leave one person out.
$ROM(j)$	Range of Motion of a joint j, computed by subtracting the extension angle from the flexion angle of such joint.
SS	Sample Synthesis
$TAM(f)$	Total Active Motion of a finger f.

We applied a depth threshold on the depth map to reduce the search space of the tracker. For each patient, we recorded movements of flexion and abduction.[2] Figs. 4 and 5 show examples of flexion and abduction movements recorded from control subjects.

Our dataset contains samples captured from 12 healthy subjects and 8 RA patients, performing flexion and abduction movements with each of their hands. For each RA patient and each hand with ulnar deviation, we obtained two flexion and two abduction sequences. The data acquired from patients is limited due to the availability of patients and occupational therapists. We performed sample synthesis (SS) in a similar way to data augmentation strategies. The goal was to evaluate the effect of noise in the hand pose measurements as well as to balance the classes in the performance assessment experiments (Section 4.3)[3].

Table 2 presents a summary of our dataset. There is a number of challenges in acquiring data from real patients of a degenerative condition that mostly affects elderly people. The challenges include matching agendas, setting up acquisition hardware in the restricted space of a busy clinic with tight schedule and getting patients who do not mind having their deformed hand captured on video. To the best of our knowledge, our dataset is the first of its kind and we made it publicly available from http://vision.ime.usp.br/~cejnog/handanalysis. This is indeed one of the main contributions of our work. The raw data captured from the sensor included RGB channels, which could potentially expose and enable the identification of patients. We reprocessed all the data to remove the RGB channels and to store the depth maps in an accessible way that is compatible with Pose-REN and does not affect the results reported.

In some of the captured sequences the patient wore an orthosis. Quoting Goia et al. [7], "orthoses are external devices applied to any part of the body to stabilize or immobilize, prevent or correct deformities, protect against injury, maximize function and reduce the pain caused by deformity". In the treatment of fingers ulnar deviation due to rheumatic arthrosis (RA), orthoses are tailor-made by therapists and act like a lever system distributing the force applied to correct the fingers ulnar deviation. The orthoses worn by the patients in our dataset were built by a team of bioengineers, mechanical engineers and occupational therapists, using a CAD system and a 3D printer, as detailed in Ref. [7].

3.2. 3D hand pose estimation

Given the unusual features of patient hands, the hand trackers that generate the best results on standard benchmarks do not necessarily perform well on our dataset. After a range of preliminary experiments, we chose Pose-REN [25] trained with the HANDS17 dataset [42], which produced the best results when applied to our data.[4] Pose-REN is based on the estimation of feature maps using Convolutional Neural Networks (CNNs). These feature maps are combined using an ensemble network, in order to generate a consistent hand pose.

The method takes as input a depth image $\mathscr{D}$ and returns as output the 3D locations $\mathscr{P} = (p_{xi}, p_{yi}, p_{zi}), i \in \{0, \cdots, N_j\}$ of the hand joints, where N_j is the number of joints of the hand model. The architecture is recurrent: the current estimate of the hand pose $\mathscr{S}_t$ is used as input to help refining it at $\mathscr{S}_{t+1}$. In the first iteration, a coarse hand pose $\mathscr{S}_0$ is estimated using a simple CNN. The network then enhances this pose in two steps: pose-guided region extraction and structured region ensemble. For a pose $\mathscr{S}_t$, in the region extraction, each point of the skeleton is projected from

[2] In the rest of this paper, we use "flexion" and "abduction" as a shorthand for "flexion/extension" and "adduction/abduction" movements.

[3] This approach has been inspired by a solution proposed in bioinformatics, where the small sample size problem is common in many situations, see Dougherty et al. [41].

[4] After the development of our experiments and the writing of this paper, the JGR-P2O [27] was published along with its source code. An evaluation of that method for RA patient image sequences is suggested as future work.

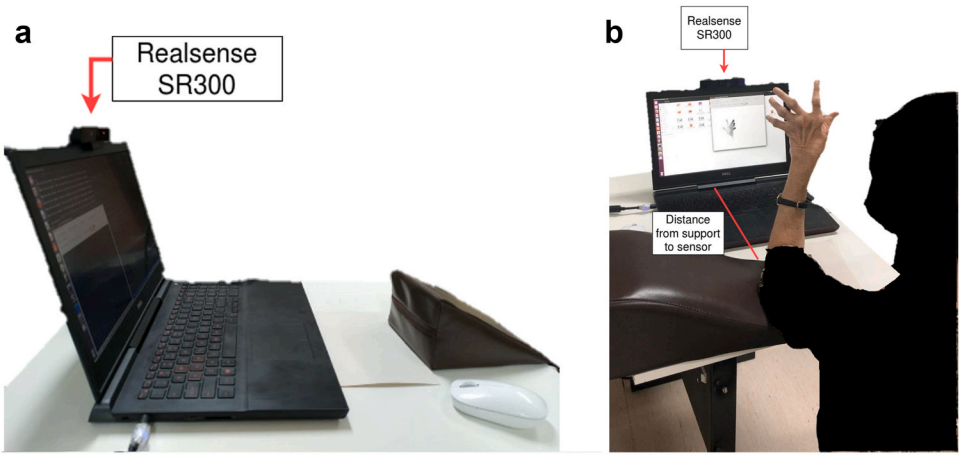

Fig. 3. Setup used for data acquisition, with the *Intel RealSense® SR300*. The SR300 has an operating range of up to 2 m. In our application the subjects were instructed to keep their hand at a distance of no more than 60 cm from camera lens.

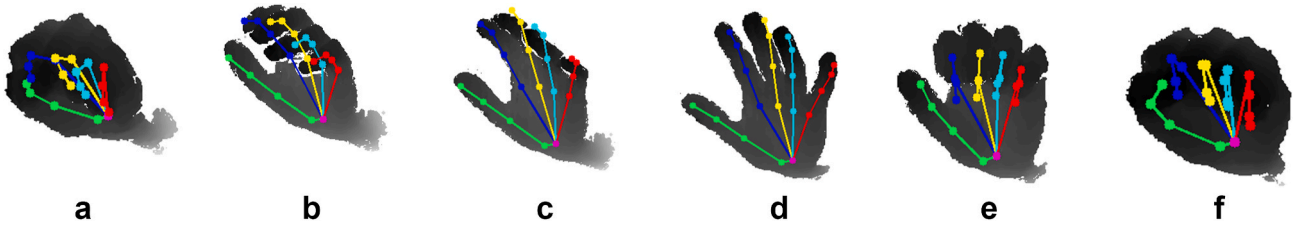

Fig. 4. Example of flexion/extension movement.

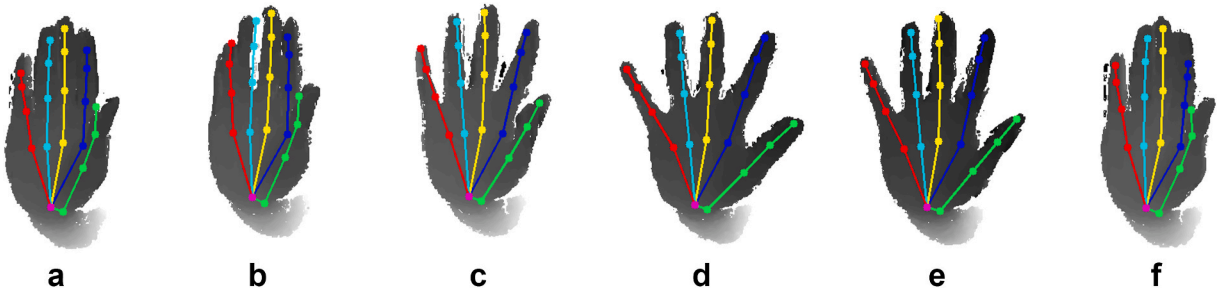

Fig. 5. Example of abduction/adduction movement.

Table 2
Summary of our dataset, available from http://vision.ime.usp.br/~cejnog/handanalysis.

Summary	
Patients with rheumatoid arthritis	**8**
Number of people in the control set	**12**
Patient Sequences	**79**
Control Sequences	**108**
Patient clips	**310**
Control clips	**581**
Total clips	**891**
Total number of frames	**85,755**
Frames used on clips	**60,192**
Percentage of frames used	**70.2%**
Size (GB)	**482**

world to pixel coordinates, and a bounding box around each joint is cropped, generating the feature regions $\mathscr{F}_i^t$. In the ensemble network, those feature regions are processed by fully-connected (fc) layers, generating for each joint j feature vectors $h_j^{l_1}$, where l_1 indicates the first fc layer. These feature vectors are integrated hierarchically according to the topology of the hand, i.e. joints that belong to the same finger are concatenated in the same vector. This vector is then fed into another fc layer, whose output is a feature vector $h_i^{l_2}$ for each finger i. The vectors of all fingers are concatenated and again fed to a fc layer, whose output is $\mathscr{S}_{t+1}$, a $3 \times N_j$ matrix of point positions in 3D. This pose is then fed into the initial layer, starting a new iteration of the network, and gradually features around the location of the fingers contribute more to the feature vectors than distant features, optimizing the output poses. This method is currently among the best performing methods in all state-of-art datasets. Its implementation is available online.[5]

The skeleton used by HANDS17 dataset has 21 points of reference: the center of the wrist (W) and for each finger k the proximal interphalangeal (PIP_k), the distal interphalangeal joints (DIP_k) and the tip (TIP_k). The exception is the thumb, which is represented by the carpometacarpal joint (CMC) and a single interphalangeal joint (IP). Fingers are represented by the respective Roman number (I–V: I for the thumb, V

[5] https://github.com/xinghaochen/Pose-REN.

for the little finger). The skeleton is represented in Fig. 6. In our pipeline, we will refer to a depth image as $D(x, y, t)$ and to the skeletons obtained by the hand pose estimation algorithm as $\overrightarrow{S}(t)$.

As a preprocessing, we perform depth filtering. As mentioned in Subsection 3.1, in most cases this is sufficient to segment the hand in the scene. Fig. 7 shows qualitative results of the Pose-REN method with patient data. In general, the method can handle the challenges of the dataset, but in some examples there are visible inaccuracies. The main challenges were the presence of orthosis, since the method was not trained to deal with hand-object interactions; situations where the arm is preeminent and not segmented by background threshold, and when the fingers are flexed in the abduction movement. From all frames captured, we utilized 70.2% in the dataset, discarding frames that are not in the predefined movement sequences and frames with inaccurate skeletons. This selection was done using a visual inspection and it shows that the pose estimation method worked well in at least 70.2% of the frames, though some frames were discarded simply because the subject was not performing the required movement.

3.3. Hand analysis

Using the skeletons $\overrightarrow{S}(t)$ obtained by the hand pose estimation method, the analysis aims to obtain measurements of flexion/extension and adduction/abduction. Such measurements are computed for each frame of all sequences obtained in the acquisition. Our ultimate goal is to estimate these angles with accuracy similar to that obtained using manual measurements with goniometers, but in a more efficient and less intrusive way.

The estimation of the flexion angles is obtained by extracting the vectors between the adjacent joints in the structure. For the finger k, the flexion angles from the joints MCP, PIP and DIP are defined respectively as:

$$\widehat{\mathrm{FMCP}}_k = \arccos\left(\overrightarrow{\mathrm{MCP}_k - W} \cdot \overrightarrow{\mathrm{PIP}_k - \mathrm{MCP}_k}\right) \tag{1}$$

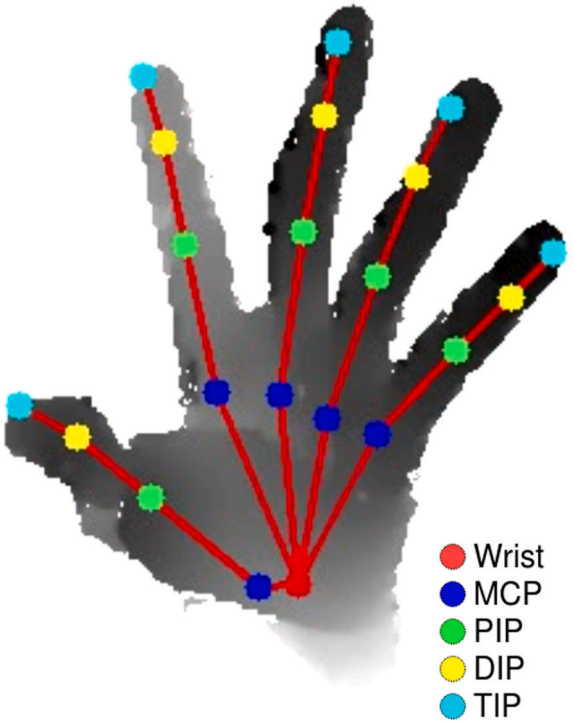

Fig. 6. Hand model used in the HANDS17 dataset.

$$\widehat{\mathrm{FPIP}}_k = \arccos\left(\overrightarrow{\mathrm{PIP}_k - \mathrm{MCP}_k} \cdot \overrightarrow{\mathrm{DIP}_k - \mathrm{PIP}_k}\right) \tag{2}$$

$$\widehat{\mathrm{FDIP}}_k = \arccos\left(\overrightarrow{\mathrm{DIP}_k - \mathrm{PIP}_k} \cdot \overrightarrow{\mathrm{TIP}_k - \mathrm{DIP}_k}\right) \tag{3}$$

For the thumb, the flexion angles of CMC and IP joints are obtained analogously.

As for abduction, there are some difficulties to compute it as an angle because the angle between two phalanx bones actually depends on two systems of joints, rather than a single joint that connects both. This makes it hard to dissociate abduction from flexion angles, particularly on hands with deformities. For this reason, it is common that occupational therapists actually measure abduction by the distance between two consecutive fingertips. In any case, we also compute the opening between the fingers, which is not a usual measurement for occupational therapy, but is straightforward and can indicate other types of patterns in a way that is invariant to the size of the hands. The opening angle is computed as the angle between the midpoint of the MCP joints of both fingers and each PIP joint.

$$\begin{aligned} \mathrm{ATIP}_k &= \left|\left|\mathrm{TIP}_{x-1} - \mathrm{TIP}_k\right|\right|_2 \\ OP_k &= \arccos\left(\overrightarrow{\mathrm{PIP}_k - mid(\mathrm{MCP}_k, \mathrm{MCP}_{x+1})} \cdot \overrightarrow{\mathrm{PIP}_{x+1} - mid(\mathrm{MCP}_k, \mathrm{MCP}_{x+1})}\right) \end{aligned} \tag{4}$$

Figs. 8 and 9 show angles $\widehat{\mathrm{FMCP}}_3$, $\widehat{\mathrm{FTIP}}_3$, $\widehat{\mathrm{FPIP}}_3$ and $\widehat{\mathrm{ATIP}}_3$ computed for all frames in a sequence obtained with a control individual and a patient, respectively. These figures also show the correspondences between given poses and maximum and minimum values on the angle graphics, thus illustrating that the method of hand pose estimation reaches consistent results for flexion movements. For the patient with ulnar deviation, the angle sequences show a higher variability, which is caused by the higher variability of the hand shapes of the patient.

Results obtained from the patient and control hands show that the Pose-REN method is able to generalize for unseen shapes, and despite the inaccuracy for unusual hand poses, the overall performance for angle detection shows that the method can be used in our pipeline.

4. Experimental methodology and results

4.1. Data description

For all sequences of movement acquired with the patients and with the control individuals we compute the flexion and abduction angles frame by frame, and manually extract the landmark frames in the beginning and in the end of each movement. These landmarks could be detected automatically using a time series analysis method, but we assume that the therapist will hit a trigger to indicate when to start measurements in the GUI to be developed as part of this project. Furthermore, we preferred to keep temporal landmark detection out of the scope of our evaluations. We will refer to the angle representation of a movement sequence as a *clip*, representing the i-th angle as $a_i(t)$.

For each detected cycle of movement, we normalize the sequences of hand points trajectories by sub-sampling them, so that all clips have the same duration (i.e., the same number of measurements). The sample representation used for the classification experiment is based on the extraction of Fourier coefficients $\mathscr{F}(i)$ for each angle a_i. The 25 first coefficients for each angle are concatenated and stored as a sample representation. The coefficients of the Fourier transform $F_{a_i}(u)$ for an angle a_i are computed by:

$$\mathscr{F}_{a_i}(u) = \frac{1}{N}\sum_{t=0}^{N-1} a_i(t) e^{\frac{-j2\pi ut}{N}}; u = 0, \cdots, 24 \tag{5}$$

Let N_a be the number of angles computed for each clip. The final representation $\mathscr{F}(u)$ of a clip is the concatenation of all Fourier descriptors of all angles.

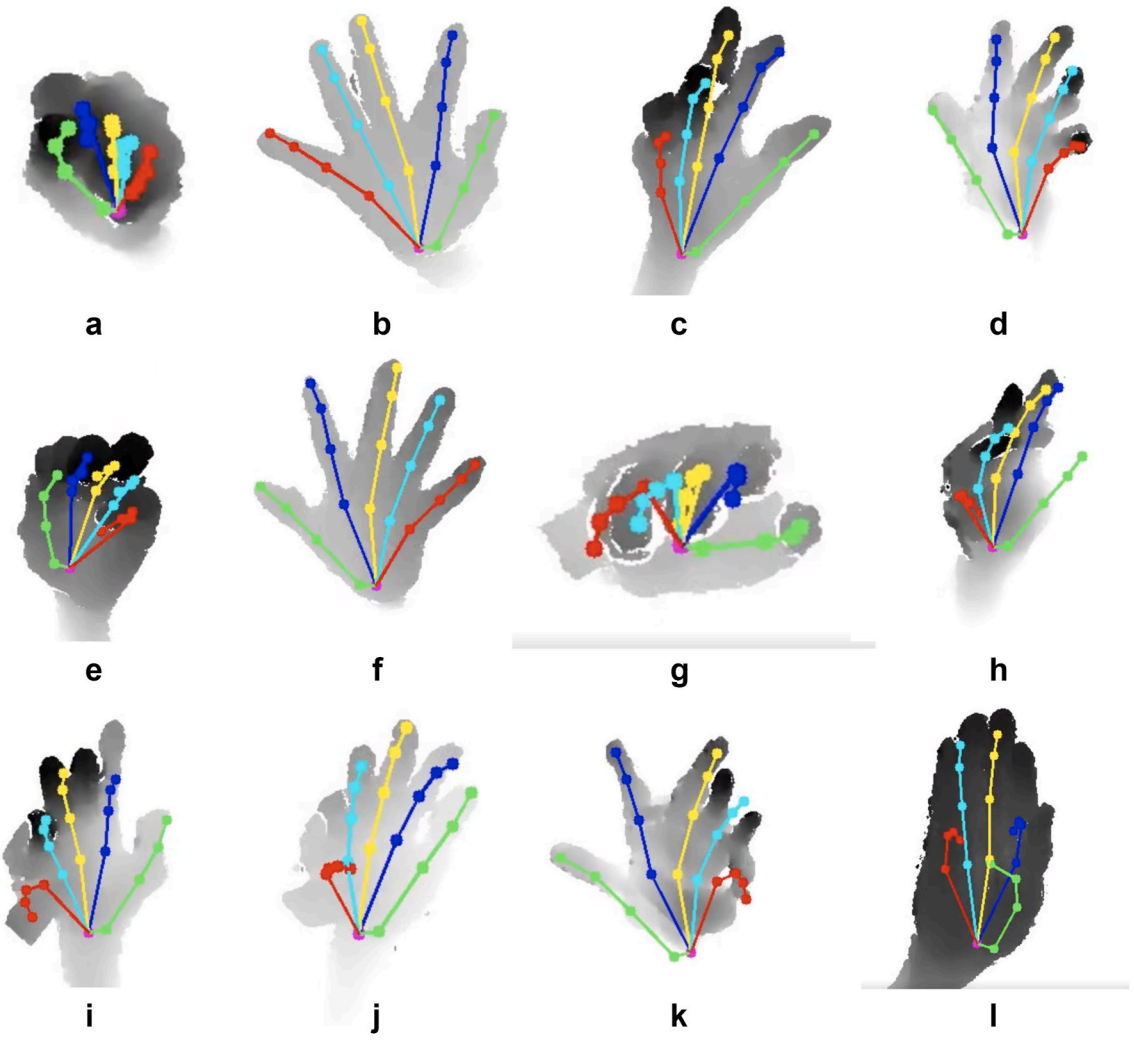

Fig. 7. Qualitative results obtained on our patients data using Pose-REN [25] trained with the HANDS17 model. Note that (h), (i), (j), (k) and (l) present some failure cases. However, those are surprisingly good results because in those images, the patient presents a high level of hand deformity and is wearing an orthosis. Results also degrade when the arm is preeminent in the image.

$$\mathscr{F} = concat(\mathscr{F}_{a_i}(u)), \tag{6}$$

for $i = 1, \cdots, N_a$ and $u = 0, \cdots, 24$. Fig. 10 shows examples of training samples, obtained after the FFT processing. Note that each sample has $25 \times N_a = 575$ dimensions.

4.2. Classification

We performed a series of experiments to classify sequences into patients or control. Since the number of samples is small, we defined three types of classification experiments: 80-20% split, leave-one-person-out, and leave-one-person-out with sample synthesis (LOO + SS).

The goal of the initial experiments was to validate the feature extraction method based on Fourier descriptors and to choose an adequate classification algorithm. To validate our method, we defined a baseline descriptor which is built by simply concatenating of the minimum and maximum value of each angle of each joint of the hand.

$$\mathscr{B} = concat(\min(a_i), \max(a_i)), \tag{7}$$

for $i = 1, \cdots, N_a$. We performed paired experiments with both baseline and Fourier descriptors, using Split and Leave-one-person-out strategies.

- Split (80–20): since the sample shuffling can affect the data distribution, we perform 10 instances of classification, each with a random split of 80% of the samples for training and the remaining for testing. We then report the mean and standard deviation of the accuracies obtained.
- Leave-one-person-out (LOO): we choose one person and take all clips from that person as the test set. Training is done with all other sequences. This test shows whether the pattern obtained from a patient or a control subject can generalize well for unseen subjects. We grouped the results in control and patient groups, showing the mean and standard deviation of the accuracy of both groups.

Additionally, different supervised classifiers have been tried in both Split 80–20 and LOO during the experiments, namely: AdaBoost, Decision Tree, Gaussian Process, Linear SVM, Naive Bayes, Nearest Neighbors, Neural Net, QDA, Random Forest and RBF SVM. Among the classifiers, the Linear SVM presented the best performance. The results of both experiments are shown in Tables 3 and 4.

The best combination of classifier and descriptor in both experiments was the Linear SVM with the Fourier descriptor, reaching an accuracy of 94.1% in the Split 80–20 experiment, and 89.6% in the leave-one-person-out experiment. It is worth mentioning that, except for some specific cases, most classifiers did not perform much worse than the SVM results here reported. Our interpretation is that the proposed hand tracking and angle measurements successfully capture the differences between control and patient movements in a robust way. Therefore, the classification task itself does not critically depend neither on the features nor on the classifier, which is a good advantage of the proposed

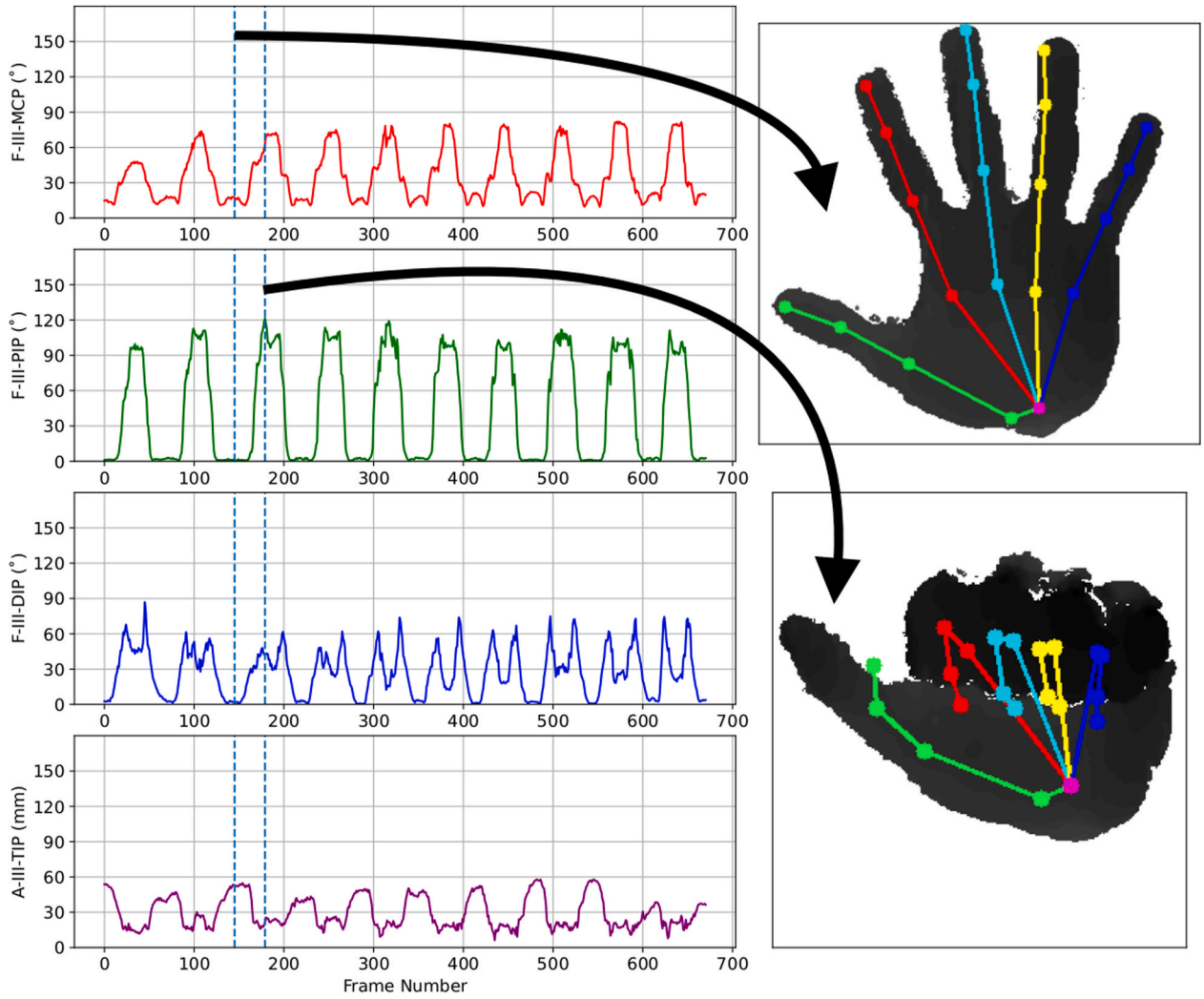

Fig. 8. Angle estimates and two correspondences to poses obtained by the pose estimation algorithm on a control individual.

framework.

The Fourier descriptor was consistently better than the baseline descriptor, with an average difference of 5%. The baseline result reached the average accuracy of 84% in the leave-one-person-out, which indicates that the minimum and maximum angles are important measurements and can be used to identify patient and control. However, the information added by Fourier descriptors is able to consistently improve the performance, working as a fine-tuned descriptor.

The high accuracy of the Linear SVM in both experiments is a good indicative, especially in the leave-one-person-out, which shows that the descriptor can be generalized for unseen subjects. For patients, the accuracy was slightly lower, which is expected as the data is more diverse, since each patient's hand is in a different stage of ulnar deviation. This higher variance in hand shapes and movement patterns creates data clusters that are more challenging for the classifier.

4.3. Data generation with sample synthesis

For the third experiment we performed sample synthesis (SS) [41] to address the inbalance between the amount of samples from patients and control. In this process, we generate synthetic data from the samples. With this, we can evaluate the results of the tracker in the presence of noise. In our work, we applied Gaussian noise for each skeleton. One could question why we have not applied standard data augmentation strategies on the depth maps, instead of injecting noise on the pose estimation data. The data augmentation strategies used in other computer vision applications usually follow two strategies: (a) RGB value perturbations (such as changing brightness, contrast, injecting Gaussian noise, etc.) and (b) homography transformations, cropping and padding. These strategies cannot naïvelly be applied to depth maps for the following reasons:

(a) The behaviour of noise in depth maps is different from that of pixel RGB values. The noise from depth sensors that are based on active infra-red patterns tend to alternate between Gaussian-like patterns and patches with unknown depth values. In fact, Yan et al. [32] have exploited the intrinsic low-rang and self-similarity property of depth images to propose a denoising method. A proper data augmentation method should start from a 3D scene model and apply a transformation that would do the inverse of what the method of Yan et al. does.

(b) Homography-based distortions would be unrealistic for our data acquisition setting and would generate on unexpected depth values. An alternative would be to geneate a 3D point cloud from each depth map, perturb the 3D position of each point and regenerate depth maps by ray tracing and interpolations.

There are works in the literature that discuss the best ways of augmenting depth maps, for problems that are different than hand pose estimation, such as depth completion [43]. However, the complexity of

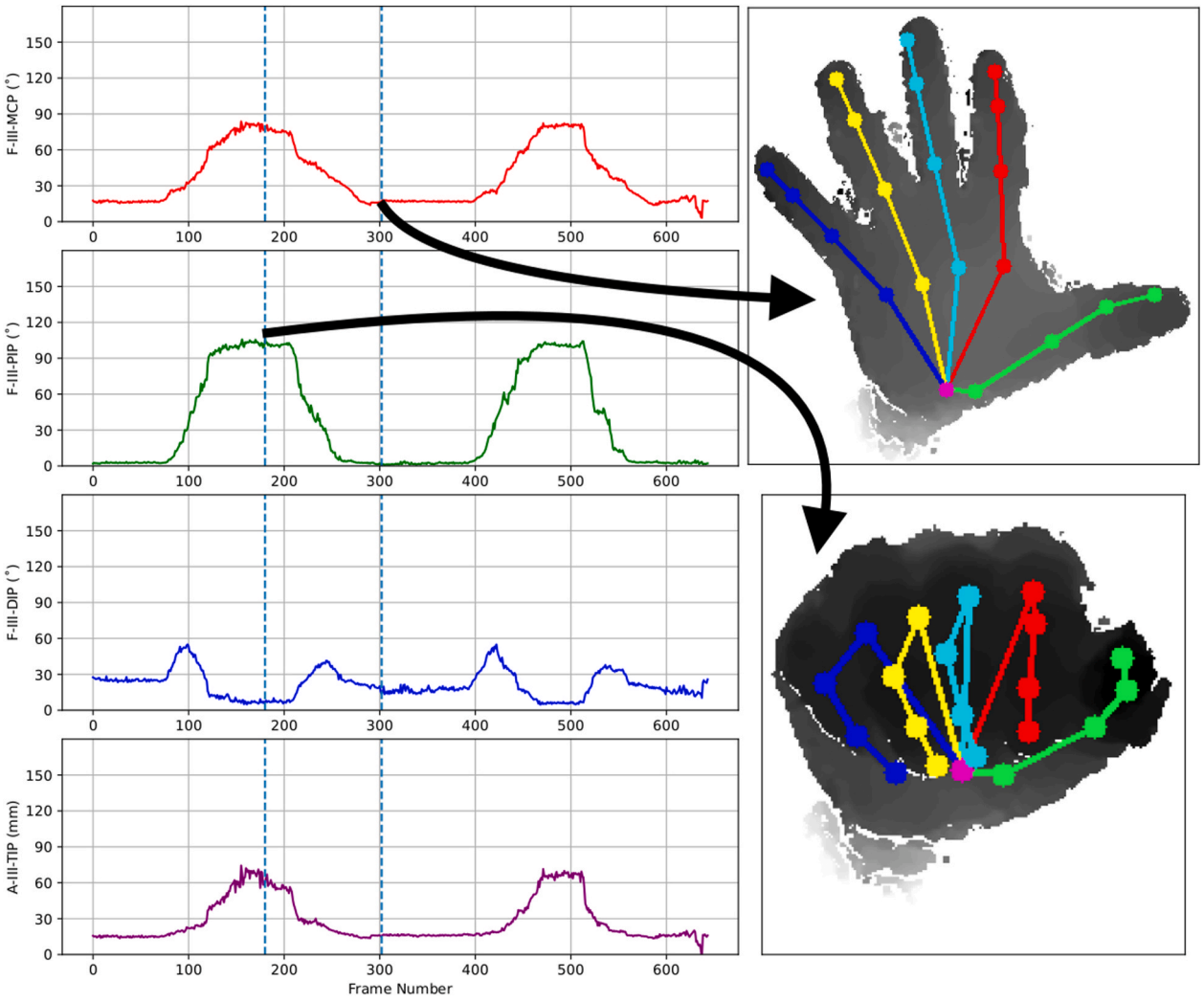

Fig. 9. Angle estimates and two correspondences to poses obtained by the pose estimation algorithm on a RA patient.

such methods would certainly slow down the training process. Furthermore, our depth maps were acquired using a real sensor in non-ideal conditions (particularly when the patients were wearing orthoses). This means that the depth maps already had a noise that is very typical of that kind of sensor and those conditions. We therefore believe that there was no need to inject further noise on depth maps to synthesize new samples. Instead, we focused our sample synthesis method on modelling potential imperfections of the hand pose estimation method (rather than on the depth maps), which is why it was more sensible to inject noise on the resulting 3D point positions. This is in line with other papers about pose estimation methods: many of them make use of skeletons and model priors for sample synthesis and for the training process [44–46].

Therefore, for a sequence

$$\overrightarrow{S}(t) = \{x_i(t), y_i(t), z_i(t)\}$$

for $i = 1, \cdots, N_j$ and $t = 1, \cdots, T$, we generate the augmented sequence

$$\overrightarrow{S}'(t) = \{x_i(t) + \mathcal{N}(0,\sigma), y_i(t) + \mathcal{N}(0,\sigma), z_i(t) + \mathcal{N}(0,\sigma)\},$$

where $\mathcal{N}(\mu,\sigma)$ represents a Gaussian function with μ mean and σ standard deviation, measured in millimeters. This procedure is applied in each frame to generate new clip samples.

In this sense, we augmented the training and the testing sets and performed the Leave-one-person-out experiment. We analyze the variations of control and patient sets and evaluate how the Gaussian noise affects the classification accuracy. For each patient hand and noise magnitude, we generated 100 augmented samples from the original sequences. The training set is composed of each execution with 100 random samples of the remaining patients, such that the number of control and patient samples is equal, and the test set is composed of all samples of the unseen patient. We repeated the leave-one-out experiment using all subjects for testing and different values of σ (1, 2 and 4). We used the Linear SVM classifier and the Fourier descriptors, which yielded the best results in previous experiments. The results are shown in Table 5.

We observe that the results obtained in the leave-one-out experiment are a more accurate representation of the classification experiment for unseen hands. The presence of augmented data lowers the accuracy for the patient set (around 75% in this case, for all values of σ), with significant values of standard deviation. This small performance loss was expected due to the variance of hand poses in the patient set, as discussed earlier. Nonetheless, the results were consistent and show that the pipeline is able to handle different levels of noise, and that the angle analysis can deal with small inconsistencies from the tracker.

5. Conclusion

We sought to evaluate the possibility of using state-of-art hand pose estimation for hand occupational therapy evaluation. We defined an

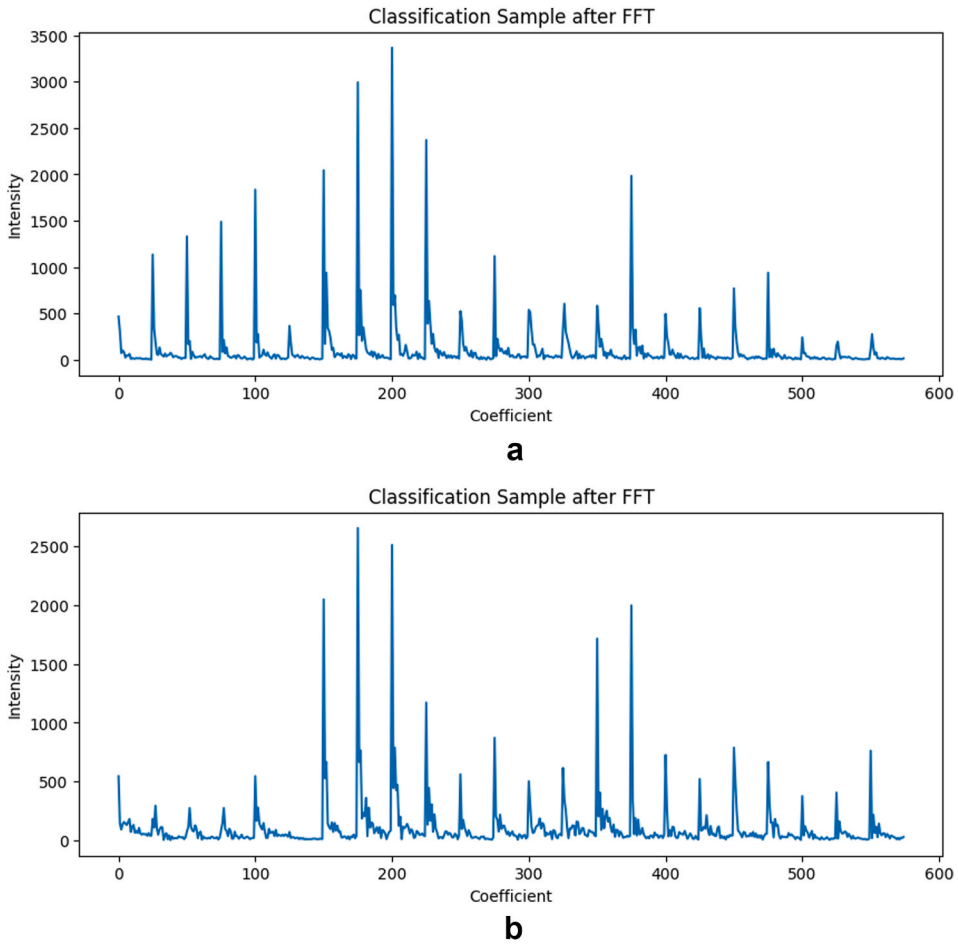

Fig. 10. Examples of Fourier descriptors, obtained through the concatenation of Fourier descriptors for each angle of a clip.

Table 3
Best performance classifiers on the Split experiment (in percentage of accuracy).

Experiment	Control (%)	Patient (%)	General (%)
Fourier Linear SVM	96.31 ± 3.07	91.97 ± 6.55	94.14± 5.56
Baseline QDA	96.66 ± 3.81	89.11 ± 6.82	92.88 ± 6.69
Fourier Nearest Neighbors	97.08 ± 3.42	86.31 ± 9.39	91.69 ± 8.88
Baseline AdaBoost	94.99 ± 4.29	83.08 ± 12.56	89.04 ± 11.11
Baseline Neural Net	88.87 ± 8.93	88.65 ± 8.03	88.76 ± 8.49
Baseline Linear SVM	92.72 ± 3.69	84.60 ± 7.51	88.66 ± 7.17
Fourier AdaBoost	95.86 ± 1.51	78.97 ± 11.65	87.41 ± 11.85
Fourier Neural Net	94.37 ± 5.88	78.50 ± 13.42	86.44 ± 13.05
Baseline Nearest Neighbors	95.30 ± 4.68	73.79 ± 13.97	84.54 ± 14.97
Baseline Random Forest	98.96 ± 1.62	65.33 ± 15.04	82.15 ± 19.93

Table 4
Best performing classifiers on the leave-one-person-out experiment.

Experiment	Control (%)	Patient (%)	General (%)
Fourier Linear SVM	94.33 ± 10.53	81.57 ± 31.34	**89.63**± 21.67
Fourier Neural Net	92.89 ± 12.01	73.11 ± 36.33	85.60 ± 25.85
Baseline AdaBoost	89.54 ± 15.54	74.07 ± 30.09	83.84 ± 23.28
Baseline Linear SVM	89.08 ± 20.29	74.57 ± 33.44	83.74 ± 26.85
Baseline Neural Net	87.35 ± 22.97	73.91 ± 35.87	82.40 ± 29.14
Fourier AdaBoost	91.92 ± 8.93	65.71 ± 34.74	82.26 ± 25.59
Fourier Decision Tree	90.76 ± 10.56	62.01 ± 28.66	80.17 ± 23.78
Baseline QDA	90.58 ± 17.43	59.71 ± 38.39	79.21 ± 30.93
Baseline Random Forest	95.82 ± 9.59	49.71 ± 40.60	78.83 ± 34.06
Fourier Nearest Neighbors	92.50 ± 16.03	53.64 ± 40.62	78.18 ± 33.49

acquisition setup with a patient set and a control set, obtaining flexion and abduction movement sequences. In case of patients, the ulnar deviation and the use of orthosis affect the acquisition and the resulting hand pose, which makes the problem even more challenging than current state-of-art hand pose estimation.

We estimated the hand pose using the Pose-REN algorithm, arguably one of the state-of-art techniques, and presented a method to convert 3D point coordinates to flexion and abduction angles. We proposed a strategy to register sequences of movements and represent cycles of movements as feature vectors based on frequency domain descriptors. This representation of the movement patterns is then used for classification, aiming to identify patterns and distinguish patients and control data.

Table 5
Linear SVM accuracy (in %) with sample synthesis using different values of σ (in mm).

Experiment	Control	Patient	General
LOO	94.66 ± 8.45	83.00 ± 31.69	90.37 ± 21.14
LOO + SS, $\sigma = 1$	88.19 ± 19.31	72.35 ± 36.87	82.35 ± 28.19
LOO + SS, $\sigma = 2$	88.61 ± 18.19	73.24 ± 36.18	82.94 ± 27.32
LOO + SS, $\sigma = 4$	86.97 ± 18.17	73.31 ± 34.96	81.93 ± 26.49

The proposed method is able to accurately estimate skeleton angles and range of motion measurements from control and Rheumatoid Arthritis (RA) patients, even with the 3D hand pose estimation algorithm being trained in a completely different dataset of healthy hand movements. Results for classification are promising, showing that a simple movement cycle is enough to distinguish patients from control.

We expect that the evolution of the state-of-art methods for hand pose estimation will lead to further advances in the analysis of patient patterns and RA diagnosis.

Declaration of competing interest

None declared.

Acknowledgments

This work received financial support of the São Paulo Research Foundation (FAPESP), grants #2015/22308–2 and #2016/13791–4. TdC thanks the support of Conselho Nacional de Pesquisa (CNPq) grant PQ 314154/2018–3. Also, we are grateful to Daniela Goia for her collaboration in the data acquisition, and Janko Calic, Maria da Graça Pimentel, Phillip Krejov and Adrian Hilton for the fruitful discussions on the project.

References

[1] Monkaresi H, Calvo RA, Yan H. A machine learning approach to improve contactless heart rate monitoring using a webcam. IEEE journal of biomedical and health informatics 2013;18:1153–60.
[2] Wang X, Gui Q, Liu B, Jin Z, Chen Y. Enabling smart personalized healthcare: a hybrid mobile-cloud approach for ECG telemonitoring. IEEE journal of biomedical and health informatics 2013;18:739–45.
[3] Adeli E, Rekik S, Park SH, Shen D. Predictive intelligence in biomedical and health informatics. IEEE Journal of Biomedical and Health Informatics 2020;24:333–5.
[4] Donate PB, de Lima KA, Peres RS, Almeida F, Fukada SY, Silva TA, Nascimento DC, Cecilio NT, Talbot J, Oliveira RD, Passos GA, Alves-Filho JC, Cunha TM, Louzada-Junior P, Liew FY, Cunha FQ. Cigarette smoke induces miR-132 in Th17 cells that enhance osteoclastogenesis in inflammatory arthritis. Proc Natl Acad Sci USA 2021;118:e2017120118. https://doi.org/10.1073/pnas.2017120118.
[5] Scott DL, Wolfe F, Huizinga TW. Rheumatoid arthritis. Lancet 2010;376:1094–108. https://doi.org/10.1016/s0140-6736(10)60826-4.
[6] da Mota LMH, Cruz BA, Brenol CV, Pereira IA, Rezende-Fronza LS, Bertolo MB, Freitas MVC, da Silva NA, Louzada-Junior P, Giorgio RDN, Lima RAC, Kairalla RA, de Melo Kawassaki A, Bernardo WM, da Rocha Castelar Pinheiro G. Guidelines for the diagnosis of rheumatoid arthritis. Rev Bras Reumatol 2013;53:141–57. https://doi.org/10.1016/S2255-5021(13)70019-1. URL: http://www.sciencedirect.com/science/article/pii/S2255502113700191.
[7] Goia DN, Fortulan CA, Purquerio BM, Elui VMC. A new concept of orthosis for correcting fingers ulnar deviation. Research on Biomedical Engineering 2017;33:50–7. URL: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2446-47402017000100050&nrm=iso.
[8] Cejnog LWX, Cesar Jr RM, de Campos TE, Elui VMC. Hand range of motion evaluation for rheumatoid arthritis patients. In: 2019 14th IEEE international conference on automatic face gesture recognition (FG 2019); 2019. p. 1–5.
[9] Marques AP. Manual de goniometria. Manole LTDA; 1997.
[10] Ellis B, Bruton A, Goddard JR. Joint angle measurement: a comparative study of the reliability of goniometry and wire tracing for the hand. Clin Rehabil 1997;11:314–20.
[11] Bruton A, Ellis B, Goddard J. Comparison of visual estimation and goniometry for assessment of metacarpophalangeal joint angle. Physiotherapy 1999;85:201–8.
[12] Carvalho RM Fd, Mazzer N, Barbieri CH, et al. Análise da confiabilidade e reprodutibilidade da goniometria em relação à fotogrametria na mão. Acta Ortopédica Bras 2012;20:139–49.
[13] Meals CG, Saunders RJ, Desale S, Means J Kenneth R. Viability of hand and wrist photogoniometry. Hand 2018;13:301–4. https://doi.org/10.1177/1558944717702471. doi:10.1177/1558944717702471. arXiv:10.1177/1558944717702471. URL:, pMID: 28391753.
[14] Tajali SB, MacDermid JC, Grewal R, Young C. Reliability and validity of electrogoniometric range of motion measurements in patients with hand and wrist limitations. Open Orthop J 2016;10:190.
[15] Pereira LC, Rwakabayiza S, Lécureux E, Jolles BM. Reliability of the knee smartphone-application goniometer in the acute orthopedic setting. J Knee Surg 2017;30:223–30.
[16] Lima L, Melo J, Fragoso T, Vieira T, Oliveira M. Fisiomotion: sistema de avaliação de pacientes portadores de artrite reumatoide usando sensor de movimentos. In: XXV congresso brasileiro de Engenharia biomédica - CBEB; 2016.
[17] Wei S-E, Ramakrishna V, Kanade T, Sheikh Y. Convolutional pose machines. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 4724–32.
[18] Newell A, Yang K, Deng J. Stacked hourglass networks for human pose estimation. In: European conference on computer vision. Springer; 2016. p. 483–99.
[19] Le V-h, Nguyen H-c. A survey on 3D hand skeleton and pose estimation by convolutional neural network. Adv Sci Technol Eng Syst J 2020;5:144–59. https://doi.org/10.25046/aj050418.
[20] Oberweger M, Lepetit V. Deepprior++: improving fast and accurate 3D hand pose estimation. In: ICCV workshop, vol. 840; 2017. p. 2.
[21] Oberweger M, Wohlhart P, Lepetit V. Hands deep in deep learning for hand pose estimation. In: Computer vision winter workshop; 2015.
[22] Zhou Y, Jiang G, Lin Y. A novel finger and hand pose estimation technique for real-time hand gesture recognition. Pattern Recogn 2016;49:102–14.
[23] Ge L, Liang H, Yuan J, Thalmann D. 3D convolutional neural networks for efficient and robust hand pose estimation from single depth images. In: Proc. CVPR; 2017.
[24] Guo H, Wang G, Chen X, Zhang C, Qiao F, Yang H. Region ensemble network: improving convolutional network for hand pose estimation. In: Image processing (ICIP), 2017 IEEE international conference on, IEEE; 2017. p. 4512–6.
[25] Chen X, Wang G, Guo H, Zhang C. Pose guided structured region ensemble network for cascaded hand pose estimation. Neurocomputing 2019. https://doi.org/10.1016/j.neucom.2018.06.097. github.com/xinghaochen/Pose-REN, Demonstration and source code available from:.
[26] Wan C, Probst T, Van Gool L, Yao A. Crossing nets: combining gans and vaes with a shared latent space for hand pose estimation. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR). IEEE; 2017.
[27] Fang L, Liu X, Liu L, Xu H, Kang W. JGR-P2O: joint graph reasoning based pixel-to-offset prediction network for 3D hand pose estimation from a single depth image. In: European conference on computer vision (ECCV); 2020. Proprint published at arXiv:2007.04646. Source code available from: https://github.com/fanglinpu/JGR-P2O.
[28] Xiong F, Zhang B, Xiao Y, Cao Z, Yu T, Zhou Tianyi J, Yuan J. A2j: anchor-to-joint regression network for 3D articulated pose estimation from a single depth image. In: Proceedings of the IEEE conference on international conference on computer vision. ICCV); 2019.
[29] Moon G, Chang JY, Lee KM. V2v-posenet: voxel-to-voxel prediction network for accurate 3D hand and human pose estimation from a single depth map. In: The IEEE conference on computer vision and pattern recognition. CVPR); 2018.
[30] Wan C, Probst T, Van Gool L, Yao A. Dense 3D regression for hand pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 5147–56.
[31] Yuan S, Ye Q, Stenger B, Jain S, Kim T-K. Bighand2. 2m benchmark: hand pose dataset and state of the art analysis. In: Computer vision and pattern recognition (CVPR), 2017 IEEE conference on. IEEE; 2017. p. 2605–13.
[32] Yan C, Li Z, Zhang Y, Liu Y, Ji X, Zhang Y. Depth image denoising using nuclear norm and learning graph model. In: ACM transactions on multimedia computing communications and applications; 2020.
[33] Stenger B, Thayananthan A, Torr PH, Cipolla R. Model-based hand tracking using a hierarchical Bayesian filter. IEEE Trans Pattern Anal Mach Intell 2006;28:1372–84.
[34] Yan C, Gong B, Wei Y, Gao Y. Deep multi-view enhancement hashing for image retrieval. In: IEEE transactions on pattern analysis and machine intelligence (PAMI); 2020.
[35] Yan C, Shao B, Zhao H, Ning R, Zhang Y, Xu F. 3d room layout estimation from a single rgb image. IEEE Trans Multimed 2020;22:3014–24. https://doi.org/10.1109/TMM.2020.2967645.
[36] Cao Z, Hidalgo Martinez G, Simon T, Wei S, Sheikh YA. OpenPose: realtime multi-person 2D pose estimation using part affinity fields. IEEE Trans Pattern Anal Mach Intell 2021;43. https://doi.org/10.1109/TPAMI.2019.2929257.
[37] Zimmermann C, Brox T. Learning to estimate 3D hand pose from single RGB images. International Conference on Computer Vision 2017;1:3.
[38] Panteleris P, Argyros A. Back to RGB: 3D tracking of hands and hand-object interactions based on short-baseline stereo. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 575–84.
[39] Mueller F, Bernard F, Sotnychenko O, Mehta D, Sridhar S, Casas D, Theobalt C. GANerated hands for real-time 3D hand tracking from monocular RGB. In: Proceedings of computer vision and pattern recognition (CVPR); 2018. URL: https://handtracker.mpi-inf.mpg.de/projects/GANeratedHands/.
[40] Panteleris P, Oikonomidis I, Argyros A. Using a single RGB frame for real time 3D hand pose estimation in the wild. In: Applications of computer vision (WACV), 2018 IEEE winter conference on. IEEE; 2018. p. 436–45.
[41] Dougherty ER, Barrera J, Brun M, Kim S, Cesar RM, Chen Y, Bittner M, Trent JM. Inference from clustering with application to gene-expression microarrays. J Comput Biol 2002;9:105–26.
[42] Yuan S, Garcia-Hernando G, Stenger B, Moon G, Yong Chang J, Mu Lee K, Molchanov P, Kautz J, Honari S, Ge L, et al. Depth-based td3D hand pose estimation: from current achievements to future goals. In: The IEEE conference on computer vision and pattern recognition. CVPR); 2018.
[43] Hammond PD. Deep synthetic noise generation for RGB-D data augmentation. Brigham Young University; 2019. Ph.D. thesis.

[44] Zhang Z, Xie S, Chen M, Zhu H. Handaugment: a simple data augmentation method for depth-based 3d hand pose estimation. 2020. arXiv preprint arXiv:2001.00702.
[45] Wu Z, Hoang D, Lin S-Y, Xie Y, Chen L, Lin Y-Y, Wang Z, Fan W. Mm-hand: 3d-aware multi-modal guided hand generation for 3d hand pose synthesis. In: Proceedings of the 28th ACM international conference on multimedia; 2020. p. 2508–16.
[46] Molchanov P, Gupta S, Kim K, Kautz J. Hand gesture recognition with 3d convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops; 2015. p. 1–7.