Injured hand therapy evaluation using hand tracking

Luciano Walenty Xavier Cejnog*, Roberto Marcondes Cesar Jr.*, Teofilo de Campos† and Valéria Meirelles Carril Elui‡
*Departamento de Ciência da Computação, IME-USP, Rua do Matão, 1010, São Paulo
Email: cejnog@ime.usp.br, cesar@ime.usp.br
†Departamento de Ciência da Computação, Universidade de Brasília
Email: teodecampos@unb.br
‡Departamento de Terapia Ocupacional, FMUSP Ribeirão Preto
Email: velui@fmrp.usp.br

***Abstract*—Hand tracking is a challenging problem in computer vision that has recently gained relevance with the development of cheap consumer-level depth cameras and virtual reality devices. The objective is to identify a hand model in a scene and track the model accurately in a sequence of frames. The main proposal of this project is the development of a framework for hand tracking and gesture analysis, using a 3D model able to express different patterns of hand pose. Methods for data acquisition, learning model parameters, hand tracking/detection in video sequences and movement analysis will be developed. Here we describe the formation of the dataset and the first tests with hand pose estimation methods. Future steps include the development of hand detection, pose estimation and tracking methods based on state-of-art, as well as the assessment of movement quantities using the joint angles from the skeletons estimated by the pose estimation methods.**

Keywords: hand tracking, hand pose estimation, computer vision, depth images.

I. Introduction

Hand gesture recognition is a challenging problem in computer vision. The objective is to accurately identify patterns of hand gestures through an input stream, usually $2D$ or $3D$ video (*Intel® RealSense, Microsoft® Kinect*). This problem has applications in robotics, activity recognition, human-computer interaction (HCI) among others. Limitations and challenges commonly cited in this problem are: the high dimensionality of the hand structure, ambiguities on the model, self-occlusions and abrupt motion [1].

Gestures can be classified as static or dynamic. Static gestures (or hand poses) are recognized in a single frame, relying on geometric features of a model. Static hand gesture recognition can be divided in hand pose estimation and hand pose recognition. Those variants differ in the sense that hand pose estimation is a regression problem, while hand pose recognition is a classification problem. Dynamic gestures are variable in time. Their characterization rely in movement information and hand tracking. This work addresses static gestures and hand pose estimation. Given a static frame, our goal is to localize the hand and estimate joint positions.

Our research is related to the project "Hand tracking for occupational therapy" (proc. FAPESP 14/50769-1) , which aims to study computer vision techniques capable of providing support to patients on hand injury recovery. In our research, we focus on rheumatoid arthritis (AR) recovery. Rheumatoid arthritis is an autoimmune chronic disease with inflammatory character, characterized by peripheral polyarthritis leading to joint deformities due to bone and cartilage erosion [2]. It affects motion functionality of the hand and the treatment requires dynamic and functional evaluations. Typically, *Disabilities of the Arm, Shoulder and Hand* (*DASH*) questionnaires are used to assess hand function during the recovery process and quantitative evaluation uses range of motion measurements. The project aims to investigate the use of computer vision techniques in order to optimize the feedback of the treatment and to produce quantitative evaluations about their movement function and evolution. Figure 1 shows an example of hand with ulnar deviation that should be handled by our framework, in contrast with a normal hand. Our framework should handle both types of hand movements.

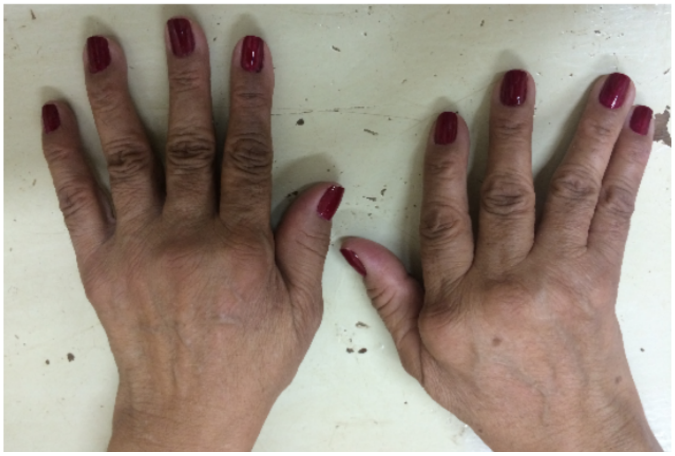


Figure 1. Example of hand with finger ulnar deviation (on the right) in contrast with a normal hand (on the left). Courtesy of Prof. Valeria Elui.

The project development should follow the pipeline proposed in Figure 2. The pipeline starts with data acquisition (RGB/RGBD), i.e. the acquisition of RGBD sequences from different patients. An initial step is the definition of a 3D hand pose estimation method, which should be applied to the captured data for the creation of a training set. The RGBD sequences with annotated joints form our dataset, which should be used for the training of a predictor capable of estimate the

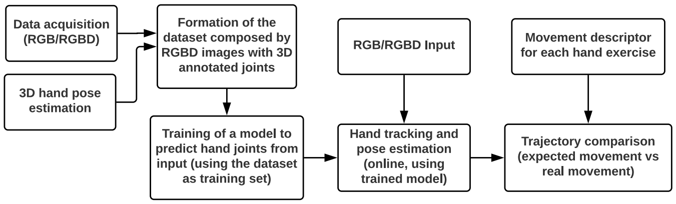

Data acquisition (RGB/RGBD)
3D hand pose estimation
Formation of the dataset composed by RGBD images with 3D annotated joints
Training of a model to predict hand joints from input (using the dataset as training set)
RGB/RGBD Input
Hand tracking and pose estimation (online, using trained model)
Movement descriptor for each hand exercise
Trajectory comparison (expected movement vs real movement)

Figure 2. Pipeline proposed.

joints for an input in similar capture conditions. To estimate accurately the hand pose from a RGB/RGBD input using the pre-trained model is the most crucial step of the project.

II. Related Works

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 CNNs [3], [4] led the community of hand pose estimation to design methods based on convolutional neural networks, reaching good results [5]–[9]. Those methods differ among themselves in the neural network architecture and type, the input image type, the hand representation used and the use of prior constraints. As an example, the *DeepPrior++* [5] uses a Residual Neural Network, a deep network whose training is based on minimizing residual weights in each layer and uses data augmentation in the training, such that realistic samples can be generated from simple geometric transformations over the original training samples. Guo *et. al.* [9] use an ensemble-based neural network, which integrates the results of different regressors in different regions of the image. Chen *et. al.* [10] compute a feature map for each joint and fuse those maps using a structured region ensemble network (named Pose-REN), reaching consistent results. Wan *et. al.* [11] propose the use of Generative Adversarial Networks (GAN) and Variational Autoencoder (VAE), two strong ideas in the recent wave of advances in machine learning. This method allows training and learning from unlabeled data.

The development of deep learning methods brought the necessity of larger datasets. As a consequence, new million-scale datasets have been made available in 2017: the BigHand2.2M [12] and First-Person Action dataset [13]. With these datasets, deep learning methods can use a much larger training set and reach better results. To consolidate the trend of using CNNs, the International Conference on Computer Vision board organized the HANDS in the million 2017 challenge on 3D pose estimation, a competition on a benchmark using the BigHand2.2M dataset. The results of this challenge were presented in the form of a survey by [14], in which design choices are discussed, as well as the corresponding evaluation results. Aspects evaluated and taken into account were:

- The nature of the input images (2D or 3D): while depth images can be seen as $2D$ points with depth, some methods perform joint detection in a $3D$ voxel grid; $3D$ volumetric representation presents high performance;
- If the method uses probability density maps (detection-based) or regresses the parameters directly from the depth image (regression-based); detection-based methods tend to outperform regression-based methods, but regression methods can reach good results using explicit spatial constraints;
- Whether the regression is hierarchical (made by subtasks, usually branches of joints are detected separately and concatenated) or holistic (the whole hand pose is regressed directly in one optimization step), and whether structural constraints and priors are incorporated in the network: the error on occluded joints is narrowed in methods with explicit modeling of structure constraints and hierarchical joints;
- Whether the training is divided in stages and one stage is used to enhance the result of the subsequent stages: cascaded methods performed better in general;
- In general, discriminative methods still generalize poorly to unseen hand shapes, and the use of models with better generative capacity can be a promising choice.

The current panorama of the area indicates that there is room for improvement on methods based on deep CNNs for depth images and that there are efforts of many research groups around the world in this direction. In parallel, new methods based on learning-based 2D joint detection and Inverse Kinematics are being proposed to estimate hand pose based exclusively on RGB image [15]–[17].

III. Proposed Approach

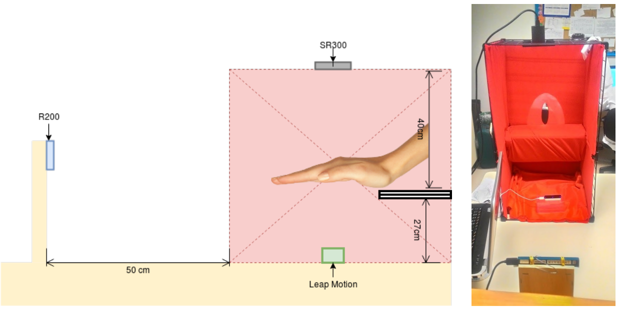
A. Dataset acquisition

As first step of the project, our goal was to acquire data from patients with rheumatoid arthritis. An acquisition setup was created using different depth sensors, in order to obtain frames in multiple views.

Initially three different sensors were studied to mount an acquisition setup: the *Intel RealSense*® *R200*, suitable for acquisitions in medium range; the *Intel RealSense*® *SR300*, that can capture points at a closer range, and the *Leap Motion*®, which generates a coarse hand tracking result, and can be used as interface for gesture recognition.

The setup was built in a way to maximize the amount of relevant information extracted from the three sensors, such that the sensors are positioned at their minimal depth range that produces stable results. This was a concern especially in the R200, since it is a medium range sensor. It was positioned to capture the hand from a frontal view with a larger distance. The SR300 captures the hand from top viewpoint, and the Leap Motion in an even shorter distance, from a bottom view. The hand is captured in an uniform background environment. Figure 3a shows a representation of the setup from a side, with the measurements of the distances used in the sensor positioning. Figure 3b shows the back view of the setup, after mounted.

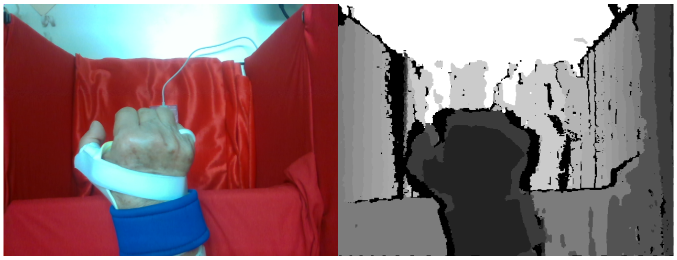
In some of the captured sequences the patient used an orthosis, a mechanical device used on the recovery in order to enhance the movement capability.



(a) Diagram of the designed setup. The sensors were positioned such that the distance to the hand is closer to the minimum range.

(b) Setup after mounted.

Figure 3. Setup used on the acquisition process.



(a) SR300 - Color

(b) SR300 - Depth

Figure 4. Example of an acquisition from a patient with orthesis, from sensors SR300 and R200.

Table I presents a summary of the dataset obtained during all visits, with the number of patients, sequences, frames and insights made in the process.

B. Preliminary results

After the data acquisition, we sought to generate initial hand pose estimation results on our dataset with two methods: Zimmerman's color-based approach [15] and Guo's Region Ensemble Network [9]. Since our data is still not annotated, we could only train the methods using standard datasets made available by the authors, doing the tests in our data. Thus, the preliminary results presented here are merely qualitative and lack statistical analysis.

1) RGB method: First we took particular interest in Zimmermann *et. al.*'s [15] approach, due to the similarity to our initial idea and the fact that the source code is available and intuitive to use. This method uses three networks in order to compute probability maps. The first network (HandSegNet) is based on the person detector provided by [3], casting the hand localization as a segmentation problem. The second network identifies 2D keypoints on the segmented region, using an architecture similar to the Pose Network (PoseNet) also presented in [3]. The following step is the application of the PosePrior network, in order to estimate the most likely 3D configuration given the 2D keypoints. This network is trained with respect to a canonical frame, and this makes the training more efficient. In our evaluation of this method no parameter tuning was done, and this might have affected significantly this output. Figure 5 presents a result of the method in one sample of our dataset.

Table I
SUMMARY OF THE OBTAINED DATA (NOT CONSIDERING CONTROL SEQUENCES).

Summary	
Patients	**9**
Sequences	**26**
Frames	**10318**
Size (MB)	**5306.4**
Sequences with orthosis	**9**

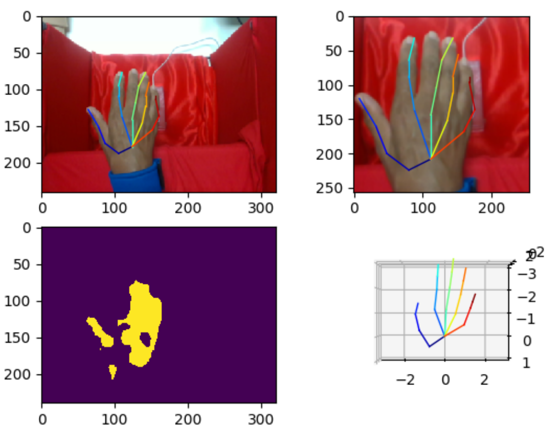


Figure 5. Sample result from Zimmermann *et. al.* in one frame of our dataset. Top left: result with respect to the whole frame, bottom left: HandSegNet result, top right: pose estimation in the cropped image after the segmentation, bottom right: 3D hand pose with respect to the global frame.

2) PoseREN: The other method we evaluated on our images was Region Ensemble Network [9]. This method computes implicit feature maps using Convolutional Neural Networks (ConvNets). The basic idea is the generation of multiple feature maps in regions of the image. Those feature maps are combined using an ensemble network, in order to generate a coherent hand pose. In our tests, we used the Realsense real-time demo provided by the authors, available on GitHub[1], in order to obtain primary results on 3D hand pose estimation. No parameter tuning was done, and the network was trained with ICVL dataset frames. This method can cope well with the orthosis, but struggles in some cases, especially when the scene is not well segmented.

3) Discussion about the results: Although promising results were obtained for simple examples and the orthosis has little influence on depth images, the method would require modifications in order to cope well with instances of our dataset. Our preliminary results presented in this chapter show that dealing with non-annotated data is a hard task with deep learning algorithms. Considering that our dataset is composed

[1] https://github.com/guohengkai/region-ensemble-network

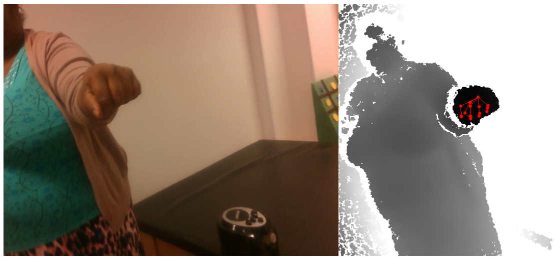


Figure 6. Result obtained for the Pose-REN method. Even with the high self-occlusion of the posterior joints the result is coarsely correct.

of different hand poses and shapes, algorithms trained in standard datasets do not reach good results. The lack of annotated data also hinders quantitative analyses with our data, which is why we are taking into account the annotation of our dataset, so that we can use it as training set to deep learning methods.

IV. Future steps

The next steps of the project are to develop a new 3D hand detection and pose estimation method based on the literature, to understand the movement according to the Occupational Therapy evaluation techniques, to measure movement from joint angle models, and to characterize different types of exercises (Flexion and Abduction) according to those movement quantities. With these three steps, we should be able to perform the evaluation of the movement patterns of the patients through 3D cameras.

Concerning the hand tracking method to be developed, one possible approach is to use machine learning methods with RGBD input, which would require data annotation, fine tuning and possibly new acquisition sessions, in order to create a training set. Another possibility is to design a method solely based on RGB data, which would make the acquisition process easier, making possible the creation of a larger dataset. The current state-of-art on the area is composed by deep learning methods based on detection of implicit features, thus it is most likely that this type of approach will be used. Since our dataset contains a large amount of unlabeled data, we can consider the possibility of developing the method using semi-supervised learning.

Therefore, we should design a method that works in a general fashion and whose training set contains hands with ulnar deviation. For this, the data annotation is necessary. Embedding our data in the learning process of the hand pose estimation algorithms would make much easier to reach good results.

V. Conclusion

The project sought to evaluate the possibility of working with affected hand poses and shapes. Despite the long road ahead, since there are still many aspects to be explorated, solid achievements such as the dataset creation and some experiments with state-of-art methods already allow us to conclude that the concept of the project is feasible.

Acknowledgment

The authors would like to thank the São Paulo Research Foundation (FAPESP), grants #2014/50769-1, #2015/22308-2 and #2016/13791-4.

Also, we are grateful to Professors Valéria Elui and Daniela Goia for their collaboration in the data acquisition, and Professors Janko Calic, Maria da Graça Pimentel, Phillip Krejov and Adrian Hilton for the fruitful discussions about the project.

References

[1] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly, "Vision-based hand pose estimation: A review," *Computer Vision and Image Understanding*, vol. 108, no. 1, pp. 52–73, 2007.

[2] P. Louzada-Junior, B. D. B. Souza, R. A. Toledo, and R. M. Ciconelli, "Análise descritiva das características demográficas e clínicas de pacientes com artrite reumatóide no estado de são paulo, brasil," *Revista Brasileira de Reumatologia*, 2007.

[3] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[4] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499.

[5] M. Oberweger and V. Lepetit, "Deepprior++: Improving fast and accurate 3d hand pose estimation," in *ICCV workshop*, vol. 840, 2017, p. 2.

[6] M. Oberweger, P. Wohlhart, and V. Lepetit, "Hands deep in deep learning for hand pose estimation," in *Computer Vision Winter Workshop*. ., 2015.

[7] Y. Zhou, G. Jiang, and Y. Lin, "A novel finger and hand pose estimation technique for real-time hand gesture recognition," *Pattern Recognition*, vol. 49, pp. 102–114, 2016.

[8] L. Ge, H. Liang, J. Yuan, and D. Thalmann, "3D convolutional neural networks for efficient and robust hand pose estimation from single depth images," in *Proc. CVPR*, 2017.

[9] H. Guo, G. Wang, X. Chen, C. Zhang, F. Qiao, and H. Yang, "Region ensemble network: Improving convolutional network for hand pose estimation," in *Image Processing (ICIP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 4512–4516.

[10] X. Chen, G. Wang, H. Guo, and C. Zhang, "Pose guided structured region ensemble network for cascaded hand pose estimation," *Neurocomputing*, 2018.

[11] C. Wan, T. Probst, L. Van Gool, and A. Yao, "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.

[12] S. Yuan, Q. Ye, B. Stenger, S. Jain, and T.-K. Kim, "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, pp. 2605–2613.

[13] G. Garcia-Hernando, S. Yuan, S. Baek, and T.-K. Kim, "First-person hand action benchmark with rgb-d videos and 3d hand pose annotations," in *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*, vol. 1, no. 2, 2018.

[14] S. Yuan, G. Garcia-Hernando, B. Stenger, G. Moon, J. Yong Chang, K. Mu Lee, P. Molchanov, J. Kautz, S. Honari, L. Ge *et al.*, "Depth-based 3d hand pose estimation: From current achievements to future goals," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[15] C. Zimmermann and T. Brox, "Learning to estimate 3d hand pose from single rgb images," in *International Conference on Computer Vision*, vol. 1, no. 2, 2017, p. 3.

[16] F. Mueller, F. Bernard, O. Sotnychenko, D. Mehta, S. Sridhar, D. Casas, and C. Theobalt, "Ganerated hands for real-time 3d hand tracking from monocular rgb," in *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, June 2018. [Online]. Available: https://handtracker.mpi-inf.mpg.de/projects/GANeratedHands/

[17] P. Panteleris, I. Oikonomidis, and A. Argyros, "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, pp. 436–445.