Real-Time Marker-Based Finger Tracking with Neural Networks

Dario Pavllo[1] Thibault Porssut *[1] Bruno Herbelin[2] Ronan Boulic[1]

[1] Immersive Interaction Group, [2] Center for Neuroprosthetics
Ecole Polytechnique Fédéral de Lausanne, Switzerland

Abstract

Hands in virtual reality applications represent our primary means for interacting with the environment. Although marker-based motion capture with inverse kinematics (IK) works for body tracking, it is less reliable for fingers often occluded when captured with cameras. Many computer vision and virtual reality applications circumvent the problem by using an additional system (e.g. inertial trackers). We explore an alternative solution that tracks hands and fingers using solely a motion capture system based on cameras and active markers with machine learning techniques. Our animation of fingers is performed by a predictive model based on neural networks, which is trained on a movements dataset acquired from several subjects with a complementary capture system (inertial). The system is as efficient as a traditional IK algorithm, provides a natural reconstruction of postures, and handles occlusions.

Index Terms: Human-centered computing—Virtual reality; Computing methodologies—Neural networks

1 Introduction

One possible way to achieve a compelling immersive virtual reality experience is to capture the user's movements using motion capture and map them on a virtual avatar. Starting from the 3D positions of markers placed on a suit, an inverse kinematics (IK) solver recovers the orientations of the avatar joints, which are in turn used for animating the avatar. The following step is the integration of the full body into the virtual reality (VR) experience, which is referred to as *embodiment*. Important limitations for the successful embodiment of the avatar are the discontinuities potentially caused by self-occlusions, and the absence of movement of the hands, both leading to a reduced sense of body ownership [3]. In such scenario, increase the number of cameras is not always feasible, and does not solve the problem entirely. Thus a real-time algorithm based on machine learning is proposed to provide an alternative to inverse kinematics, and addresses the issue of occlusions through a prediction system.

1.1 Related Work

The most common approach for correcting occlusions is to use interpolation algorithms. In this regard, some interpolation techniques have been specifically designed for human body tracking and skeleton animation [9]. However, interpolation algorithms require knowledge of both past and future data, and therefore can only be applied in post-processing. A. Aristidou et al. [2] proposed an approach based on Kalman filters for estimating the positions of occluded markers in real time. Their method does not require any prior knowledge of the skeleton, but assumes that the distance between neighbouring markers is approximately constant, which cannot be verified in many hand postures. Piazza et al. [7] developed a real-time extrapolation algorithm which assumes that motion can be either linear, circular, or a combination of both. As before, it does not rely on a predefined skeleton model, and, according to the authors, it can reliably predict the positions of occluded markers for up to two seconds. Finally, a large portion of research in this field exploits the assumption that an underlying skeleton model is available, thereby allowing the algorithm to put some constraints on the solution. One approach was proposed by Herda et al. [4], and is specifically targeted at passive motion capture systems, since they suffer from marker merging and tagging problems. This however does not address the problem of predicting the marker positions during occlusions. A different solution for capturing movement is to perform body or hand reconstruction through images and depth cameras. These approaches leverage computer vision and machine learning algorithms [5], and aim at providing a cheaper and consumer-ready alternative to complete motion capture systems. For finger tracking, one such system is the Leap Motion controller; its usage is however still quite limited in terms of range of motion and is optimized for a user facing the device, not for standing and moving.

As for machine learning, current approaches for handling occlusions are restricted to the sub-problem of real-time posture and gesture recognition [6]. In our case, we do not perform such classification tasks, and our aim is rather to achieve a complete reconstruction of the hand posture.

1.2 Contribution

Our work focuses on motion capture with active markers and proposes a machine learning-based alternative to analytical IK algorithms, as well as a method for correcting occlusions. Machine learning has already been successfully applied to inverse kinematics in a computer vision context [10] and in a traditional setting (e.g. robot arm) [1], but we aim to obtain a data-driven reconstruction of realistic poses instead of solving a constrained optimization problem. Our method predicts the most likely posture by exploiting current data (non-occluded markers) and prior information inferred from a dataset.

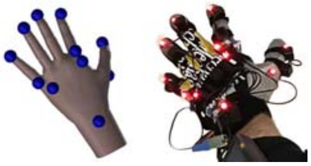

Figure 1: The virtual hand (left) and the mocap glove (right).

2 Methodology

2.1 Data Acquisition

ImpulseX2 motion capture system (optical active makers by Phasespace) and Perception Neuron gloves (low-cost inertial measurement units by Noitom) are used to acquire the data. The two were combined on the custom glove shown in Fig. 1 and used simultaneously for computing the ground truth through a sensor fusion algorithm. The resulting dataset contains the absolute positions (i.e. 3D points) of the markers (the inputs of the system) as well as the angles of each joint in the hand (the outputs of the system).The data are recorded

[*] e-mail: thibault.porssut@epfl.ch

2018 IEEE Conference on Virtual Reality and 3D User Interfaces
18-22 March, Reutlingen, Germany
978-1-5386-3365-6/18/$31.00 ©2018 IEEE

from 4 subjects with different hand sizes, who were instructed to perform various gestures, including the most problematic ones (e.g. finger crossing). The dataset is split into 3 parts: training set, validation set, test set, separating different recording sessions in order to avoid correlated samples.The training set consists of ≈30 minutes of data recorded at 60 FPS. The fingers become occluded very easily, and most occlusions involve few markers. Thus, the probability that multiple markers are occluded at once is low. Moreover, the duration of an occlusion follows a heavy-tailed distribution (90% of occlusions last less than 0.36 s).

2.2 Machine Learning Model

Our prediction model consists of a modular two-stage pipeline, which is depicted in Fig. 2. The input points are transformed into object space by applying a rigid motion transformation towards the hand template shown in Fig. 1 (left). The predicted output is then put back into world space. Furthermore, we enforce temporal consistency through an offset in object space that is computed every time an occlusion occurs.

Marker Predictor This first stage is a 5-layer autoencoder neural network that predicts the positions of the occluded markers. All layers use ReLU activation functions, except for the output layer, which uses a linear activation function. All hyperparameters were chosen according to the cross-validation results. Moreover, we use Dropout [8] in the input layer to make the network learn how to handle missing values, i.e. occluded markers.

Joint Predictor The second stage is a 5-layer dense feed-forward neural network that reconstructs the angles of all joints (similarly to an IK solver), assuming that there is no occlusion. As before, all layers except the last one use ReLU activation functions. Moreover, we use Dropout in the hidden layers to avoid overfitting.

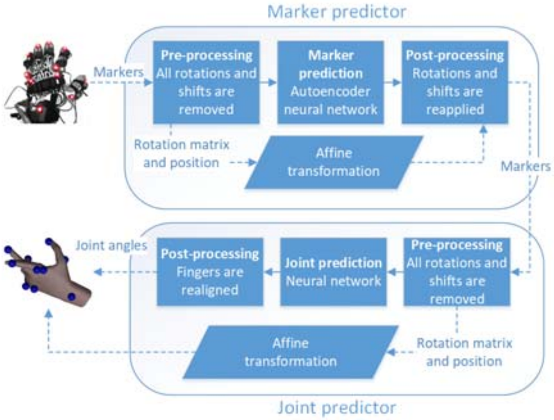

Figure 2: Full prediction pipeline. Each stage can be replaced or used separately.

2.3 Training and Implementation

The models were trained with Keras, using Theano as backend (a deep learning library written in Python). For both models, we optimized the mean squared error (MSE) loss function using the Adam optimizer. Samples with missing values were excluded from the training set, as a reliable ground truth cannot be computed for those. Occlusions are simulated using Dropout while training the model (a process known as data augmentation).

3 Results

3.1 Reconstruction Error

On the test set, the joint predictor produces a RMSE (root mean squared error) of 5.55 degrees and a MAE (mean absolute error) of 3.82 degrees, which are further reduced through a post-processing algorithm that aligns the directions of the fingers toward the markers. The marker predictor, in a pessimistic simulated scenario where occlusions have a duration of 0 to 3 seconds (uniformly distributed), yields a RMSE of 0.38 cm and a MAE of 0.15 cm. By contrast, a simple algorithm that keeps the last known position of the marker would lead to a RMSE of 12.88 cm and a MAE of 7.07 cm.

3.2 Performance

Our reference implementation is based on Unity engine. Running the entire pipeline on an Intel Core i5-4460 CPU (at 3.2 GHz) requires less than 1.2 milliseconds (≈ 833 frames per second).

4 Conclusions and Future Work

Our system provides a natural reconstruction of the hands in most real-case scenarios. Our data-driven approach to inverse kinematics does not require defining a set of rules or constraints, as these are learned automatically from the data. Occlusions are corrected with good accuracy in most cases, and with minimal latency.

In the future, we may use Recurrent neural networks to handle discontinuities without explicit corrections like LSTM (long short-term memory) which can keep track of long contexts. Furthermore, additional evaluation steps (including a comparison with other methods) would highlight the limitations and advantages of our model.

Acknowledgments

Suppported by the SNF, ”Embodied Interaction” N°200020_159968

References

[1] A. R. Almusawi, L. C. Dülger, and S. Kapucu. A new artificial neural network approach in solving inverse kinematics of robotic arm (denso vp6242). *Computational intelligence and neuroscience*, 2016.

[2] A. Aristidou, J. Cameron, and J. Lasenby. Real-Time Estimation of Missing Markers in Human Motion Capture. In *2008 2nd International Conference on Bioinformatics and Biomedical Engineering*, pp. 1343–1346. IEEE, may 2008. doi: 10.1109/ICBBE.2008.665

[3] S. Bovet, H. Galvan Debarba, B. Herbelin, E. Molla, and R. Boulic. The Critical Role of Self-Contact for Embodiment in Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics*, (part of IEEE VR conference), 2018.

[4] L. Herda, P. Fua, R. Plänkers, R. Boulic, and D. Thalmann. Skeleton-based motion capture for robust reconstruction of human motion. In *Proceedings of the Computer Animation*, CA ’00, pp. 77–. IEEE Computer Society, Washington, DC, USA, 2000.

[5] T. B. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2):90 – 126, 2006.

[6] C. Mousas and C.-N. Anagnostopoulos. Real-time performance-driven finger motion synthesis. *Computers & Graphics*, 65(Supplement C):1 – 11, 2017. doi: 10.1016/j.cag.2017.03.001

[7] T. Piazza, J. Lundström, A. Kunz, and M. Fjeld. Predicting Missing Markers in Real-Time Optical Motion Capture. *LNCS*, 5903:125–136, 2009.

[8] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014. doi: 10.1214/12-AOS1000

[9] D. Wiley and J. Hahn. Interpolation synthesis of articulated figure motion. *IEEE Computer Graphics and Applications*, 17(6):39–45, 1997. doi: 10.1109/38.626968

[10] X. Zhou, Q. Wan, W. Zhang, X. Xue, and Y. Wei. Model-based deep hand pose estimation. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI’16, pp. 2421–2427. AAAI Press, 2016.