Learning Three-dimensional Skeleton Data from Sign Language Video

HEIKE BROCK, Honda Research Institute, Japan

FELIX LAW, University of British Columbia, Canada

KAZUHIRO NAKADAI, Honda Research Institute, Japan

YUJI NAGASHIMA, Kougakuin University, Japan

Data for sign language research is often difficult and costly to acquire. We therefore present a novel pipeline able to generate motion three-dimensional (3D) skeleton data from single-camera sign language videos only. First, three recurrent neural networks are learned to infer the three-dimensional position data of body, face, and finger joints for a high resolution of the signer's skeleton. Subsequently, the angular displacements of all joints over time are estimated using inverse kinematics and mapped to a virtual sign avatar for animation. Last, the generated data are evaluated in detail, including a sign language recognition and sign language synthesis scenario. Utilizing a neural word classifier trained on real motion capture data, we reliably classify word segments built from our newly generated position data with similar accuracy as motion capture data (absolute difference 3.8%). Furthermore, qualitative evaluation of sign animations shows that the avatar performs natural movements that are comprehensible and resemble animations created with original motion capture data.

CCS Concepts: • **Computing methodologies** → **Neural networks**; *Animation*; *Motion processing*; *Tracking*; • **Human-centered computing** → **Accessibility systems and tools**;

Additional Key Words and Phrases: 3D pose estimation, recurrent neural networks, sign language synthesis, sign language recognition, data augmentation

ACM Reference format:

Heike Brock, Felix Law, Kazuhiro Nakadai, and Yuji Nagashima. 2020. Learning Three-dimensional Skeleton Data from Sign Language Video. *ACM Trans. Intell. Syst. Technol.* 11, 3, Article 30 (April 2020), 24 pages. <https://doi.org/10.1145/3377552>

1 INTRODUCTION

Labeled data constitutes a decisive resource for the development of highly accurate neural networks in artificial-intelligence driven systems of the text, image, or audio domain [17, 31, 51]. However, such extensive datasets can hardly be found for more specialized or multi-modal machine

The authors thank Michael Gienger and his team at Honda Research Institute Europe for the support regarding the use of the IK computation and the corresponding robotics library.

Authors' addresses: H. Brock, Honda Research Institute, 8-1 Honcho, Wako-shi, Saitama, Japan, 351-0188; email: h.brock@jp.honda-ri.com; F. Law, University of British Columbia, 2329 West Mall, Vancouver, Canada, BC V6T 1Z4; emails: felix.law@jp.honda-ri.com, law.yk.felix@gmail.com; K. Nakadai, Honda Research Institute, 8-1 Honcho, Wako-shi, Saitama, Japan, 351-0188; email: nakadai@jp.honda-ri.com; Y. Nagashima, Kougakuin University, 2665-1 Nakanomachi, Hachioji-shi, Tokyo, Japan, 192-0015; email: nagasima@cc.kogakuin.ac.jp.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

2157-6904/2020/04-ART30 \$15.00

<https://doi.org/10.1145/3377552>

Fig. 1. Our proposed system for use with sign language videos as found on video sharing sites. Obtained motion information could be used as additional training data for both sign recognition and sign avatar synthesis in the future.

learning tasks. Here, multiple data input and output channels easily impose additional data collection and annotation requirements that consume high amounts of both financial and human efforts. In particular, once highly accurate and occlusion-free three-dimensional human movement data are desired, the deployment of costly optical motion capture systems often becomes inevitable. In this work, we address the need of facilitated data acquisition for Sign Language (SL) translation research (Figure 1), which often requires the underlying training data to include detailed information on face and finger movement. This scenario both includes the understanding of signed movement as a sequence of words (referred to as Sign Language Recognition, SLR) and the generation of new signed expression from a word input sequence (referred to as Sign Language Synthesis, SLS).

For SLR, an increasing number of neural network variations have been reported that examine sign translation within a continuous expression of consecutive signed items. These train machine knowledge in a purely data-driven, end-to-end fashion and are mostly based on a public dataset comprised of videos of signed weather forecasts from German television broadcasts [15]. Consequently, they translate image sequences into an output sequence of gloss annotations in German SL. Enhanced network architectures drastically improved Word Error Rates (WER) for sentence translation for the given dataset, for example, from 42.2% [6] to 38.8% [30] and 38.3% [20] WER. By utilizing continuous training data instead of isolated single signs, the given architectures constitute a large step toward the development of future functional translation devices. However, due to the lack of further similarly large datasets, these architectures could neither be tested under different SLs, nor could modifications be explored for use with different data modalities, such as skeletal data.

SLS generally aims to animate virtual sign avatars in a natural and comprehensible way, so that information can be provided to deaf or hard of hearing individuals in the most accessible manner. For this, one common strategy is to combine single-word motion data segments into full sentence animations by concatenation and inter-word interpolation [13, 34, 38]. However, these animations often lack comprehensibility and user acceptance [28], mainly because it is very difficult to represent important linguistic features such as non-manual expression and directional changes sufficiently well [23, 26]. Additionally, avatar movement can easily look unnatural and artificial [22, 29]. To try to address these challenges, a first sequence-to-sequence network for direct translation of word input sequences into movement data has been reported recently. To date, it clearly suffers from a lack of training data and is not able to generate meaningful multi-words movement sequences [2].

In conclusion, both SLS and SLR research could largely benefit from methods that provide new or additional training data in a ubiquitous and accessible fashion. For this reason, we propose a machine learning-driven pipeline that generates three-dimensional (3D) skeleton data from video

input, which could then be used for implementation and training of translation systems. Here, we aim to capture the full range of information conveyed within a signed expression, meaning not only the displacements of a signer's larger body joints but also fine-grained movements of the face and fingers as necessary for animation of a virtual avatar. In particular, once our pipeline is combined with existing methodologies for the identification of relevant video content from video sharing sites [42, 47], new possibilities for the acquisition of extended machine learning datasets should then be created.

The idea of augmenting data collections for improved accuracy of SL translation systems is not new and has been particularly explored for SLR scenarios. For example, in 2007, Farhadi and White investigated the potential of transfer learning utilizing virtual avatars [14]. However, data was commonly not provided under high-resolution skeletal structures as necessary for animation of a virtual avatar in a SLS setting. Oppositely, the proposed system enables us to not only perform SLR but also to drive the movement of a virtual avatar comprised of 107 body, face, and finger joints. To the best of our knowledge, it is the first time that a neural network architecture for the inference of such high-resolution human skeletal data from single-camera video is proposed. It is also the first time a state-of-the-art pose estimator is evaluated within the context of SL translation, making our work a unique contribution to the field. In concrete, the contributions of our work are as follows: (i) we propose a neural network combination for the reliable estimation of full-body 3D joint trajectories from single-camera video specific to the characteristics of signing, (ii) we modify and extend robotic computations to determine joint angles of a virtual sign avatar from the estimated joint positions, (iii) we combine the previous novelties into a functional pipeline for the synthesis of artificial data from SL videos, and (iv) we evaluate the pipeline with respect to its actual intended application scenarios.

The rest of this article is organized as follows. Section 2 reviews related work and puts them into context with the SL target application. Section 3 describes the basic methodology of the chosen approach. In Section 4, we describe the first part of our proposed system solution, a multi-RNN architecture for generation of full-body joint position data including face and finger data. The second part of the system based on kinematic computation for avatar generation is described in Section 5. Section 6 reports experiments for both SLS and SLR contexts that are discussed in more detail in Section 7. We conclude our work in Section 8.

2 RELATED WORKS

The estimation of human pose and shape from video constitutes a popular unsolved research problem in the computer vision field [41, 46]. Given sufficiently accurate and dense-labeled data, it can be formulated as a standard supervised learning problem. Multiple collections of synchronized video and motion capture data have been created throughout the years for this purpose, for example the Human3.6M data base [24]. However, these benchmark datasets commonly do not support the tracking of finger or facial movement. Correspondingly, only very few works have been reported that aim to estimate the movement of these fine-grained body parts. One example is the face reconstruction network proposed by Jackson et al. [25]. As opposed to our work, however, the authors train a network that estimates 3D facial geometry. In the following, we will therefore discuss only these systems that are closely related to our problem, the estimation of body joint displacements. To avoid ambiguity, let us refer to pose estimation as the inference of full body joints including face and finger joints. Conversely, the estimation of major body parts (torso, legs, arms, and head) in 3D space shall be referred to as stick figure.

Several neural network architectures for stick figure estimation from video have been presented within the last years. A simple strategy is to train a supervised regression network for direct inference of joint locations [35]. More recent works propose to combine 2D and 3D information to

Table 1. Comparison between the Outputs and Applied Evaluation Strategies of Selected Pose Estimation Methods and the Proposed System

Ref. Architecture	Stick Fig. Estim.	Finger Estim.	Face Estim.	Evaluation Strategy
Tekin et al. [52]	Yes	No	No	Sequence similarity
Martinez et al. [37]	Yes	No	No	Sequence similarity
OpenPose [7] 2D	Yes	Yes	Yes	Sequence similarity
OpenPose [7] 3D	Yes (multi-cam)	No	No	Sequence similarity
Pavlakos et al. [45]	Yes	No	No	Sequence similarity
Mehta et al. [39]	Yes	No	No	Sequence similarity
MediaPipe [36] 2D	No	Yes	Yes	Sequence similarity
Proposed system	Yes	Yes	Yes	Sequence similarity, SLR & SLS use case

improve accuracy and robustness, such as the network architecture by Tekin et al. [52]. One of the most popular stick figure estimation models is OpenPose [7, 8]. However, for a single-camera input video as given in our target scenario, OpenPose provides 2D positions only. The 3d-pose-baseline model by Martinez et al. [37] utilizes the 2D joint position data obtained by OpenPose and estimates the missing third dimension by a simple feed-forward network to produce reasonable moving stick figure outputs. Pavlakos et al. [45] combine ConvNets with a volumetric stacked hourglass approach to achieve an enhanced 3D stick figure estimate. Last, the VNect architecture by Mehta et al. [39] utilizes kinematic information of the human skeleton to make their stick figure estimates more robust.

Beyond stick figures estimation, two major frameworks offer possibilities to track hand and facial joint positions: OpenPose [49] and Google’s MediaPipe [36]. However, for both networks the estimation of hand and face joints is restricted to image dimension. As a consequence, existing systems cannot be deemed sufficient for SL research that largely relies on finger movements and facial expressions. Our proposed solution consequently should hence fill the gap and provide a specialized framework for the SL context (Table 1). For training and evaluation of our system, we utilize a large collection of video and high-resolution motion capture data. Although containing multiple hours of signing, previous investigations clearly showed that the current data itself is not enough to learn and implement a direct SL translation system [2], highlighting the need of the proposed training data synthesis pipeline.

2.1 Pose Estimation and SL Data

The tracking of signing constitutes a special sub-task of human pose estimation that combines hand, finger, and face movements. As movement mainly occurs in the upper half of a signer’s body, input data differs considerably from common full-body pose tracking video data with respect to image section (upper body close-up), range of movement (on place) and ratio between moving body parts (static hips and flowing wrists). For signing, smaller positional estimation errors might furthermore already have a big impact and change the semantic content or meaning of the generated data. Earlier works of SLR mainly track SL movements with the help of basic image and signal processing methods. For example, the works of Coogan et al. [11] and its successor Han et al. [18] first extract skin objects utilizing a skin color model that are then tracked with the help of a Kalman Filter to handle occlusions between skin objects. A skin color model is also learned in the work of Charles et al. [9] to extract skin objects from a background segmented sign video sequence. Here, joints positions are then estimated with the help of a Random Forest Regressor. Last, Buehler et al. [5] fit an upper body robotic model that is subsequently applied to robustify

Fig. 2. The avatar skeleton contains 107 joints in total with a dense finger and face structure, making the number of necessary joint position estimations much larger than estimations of previous tracking systems.

a given pose detection. Common to all of the previous methods is that they focus on major body landmarks, such as elbow, wrist, and shoulder, or simple features such as hand shape and hand position. Making use of more recent methods, Gattupalli et al. [16] transfer learn two deep learning neural approaches to upper body sign language videos but similarly also only track the positions of six basic stick figure joints. In this work, we aim to bring both accuracy and resolution to a new level by fitting data to the dense skeletal hierarchy of a high-fidelity virtual avatar (Figure 2).

It appears reasonable to make use of any of the previous machine learning models trained on extensive data bases, such as the 3.6M data base, for the development of a robust data synthesis system. However, stick figure estimation models trained on large body movements might not be accurate enough for our target SL applications. We therefore first investigate whether existing models could be used in the given application setting and utilize a pre-trained pose estimation model to obtain a 3D stick figure estimate for our data generation pipeline. To illustrate our results, we will focus here on one sample model that is publicly available and very simple and free to use, namely the 3d-pose-baseline model developed by Martinez et al. [37]. Similar observations are also made when employing other models, including one commercially available system.

Following the common procedure of the 3d-pose-baseline model, we first extract the 2D joint positions of the human figure with OpenPose, and then feed the joint trajectories as input to the 3d-pose-baseline model to infer the corresponding 3D joint positions. In a subsequent visual display of the obtained pose predictions, it is easy to observe that the 3D stick figure output is fairly noisy and erroneous (Figure 3). For example, the general pose shows a considerable, permanent forward bend of the spine and the shoulder regions. Furthermore, the whole stick figure may twitch up and down or the stick figure may stretch and bend its limbs in inhuman ways.

To determine the qualitative error of the observed deformations, we calculate the jointwise Inter-Quartile Range (IQR) over the whole dataset. According to Tukey's fences ($1.5 \times \text{IQR}$ rule), 71.81% of all takes then contain outliers within the positional range of at least one joint of the stick figure, suggesting that the majority of all output estimates contains considerable errors. As quantitative error measure, we compare the estimated 3D positions to the motion capture ground truth. For

Fig. 3. Screenshots of 3D stick figure poses estimated with an existing network architecture. Every stick figure represents a randomly chosen sample frame taken from four different video footages. Errors and deformations of the human skeleton are clearly visible. Changes in size and position of the stick figure are caused by variations and noise in the network output.

this, we rotate all 3D points so that the line segment connecting left and right hip joints is aligned to the y -axis of the motion capture coordinate system. First, we determine the quaternion q_r that represents the skeleton's rotation around the longitudinal axis with the half-angle formulae [32]. Input to this computation are the normal vector $\vec{n}_y$ aligned with the y -axis and the line segment vector $\vec{v}_{h,r;h,l}$ defined by the 3D points $P_{h,r}$ and $P_{h,l}$ of the left and right hip, respectively. We then apply q_r to the 3D points P_{jk} of all k skeletal joints j as

$$P'_{jk} = q_r P_{jk} q_r^{-1} \quad (1)$$

to obtain the rotated 3D points P'_{jk} . We furthermore translate, normalize, and rescale all P'_{jk} to correlate the inferred figure's root position and joint ranges to the motion capture coordinate system and the reference stick figure's joint ranges. The normalization and rescaling of the estimated data helps to reduce the deformation errors and brings the stick figure into a more natural shape. As video and motion capture system were synchronized during data collection, the comparison can then be formulated as a simple framewise measure of vector similarity. We use the average L_1 distance between the estimated sequence s of i final joint positions P'_{jk} and the ground-truth sequence g of i absolute joint positions P_{jk}^* as a measure of the total deviation:

$$d = \frac{1}{m \cdot i} \sum_{m,i} |s_{m,i} - g_{m,i}|, \quad (2)$$

whereas m designates all captures within the dataset and i every frame within a data capture. As a result, we obtain an overall error of $d_{3dBL} = 3.609$ cm with standard deviation $\sigma_{3dBL} = 3.137$ cm.

The determined estimation error is in line with the accuracies reported by the authors. However, our analysis indicates a high risk of temporary noise and estimation errors, which might not be sufficient for the intended target application. Evaluation furthermore shows that output estimates tend to vary between different signer videos and that a reasonable pose estimate can only be obtained by applying multiple normalization and rescaling steps. The observed deformations are likely caused by the pre-trained model that does not generalize well to the specifics of SL videos. Consequently, we decide to learn a novel pose estimator for all three sub-parts stick figure, face and hands of a signing video.

Fig. 4. Overview of the system flow. Unknown video data content can be fed to the RNNs as test data.

3 METHODOLOGY

On the whole, our system is defined by two main steps, namely the estimation of global, full body joint positions from a sign video input and the computation of a virtual sign avatar's framewise movement for subsequent animation (Figure 4). To learn a specialized 3D pose estimator, we utilize the 2D joint positions of stick figure, finger, and face of a signer provided by OpenPose and feed all relevant 2D joint positions into three separate Recurrent Neural Networks (RNNs). The RNNs refine the initial 2D joint positions and simultaneously infer their missing depth information independently for all three pose sub-parts. The second main step of the video processing pipeline is based on the combined output positions of all RNNs. Following the skeletal hierarchy definitions of a chosen avatar, we determine angular and translational displacements of its respective joints by Inverse Kinematic (IK) calculus [12]. These displacements can then be used to drive a virtual avatar. For this, we combine the obtained joint rotations and translations into a motion capture BVH (Biovision Hierarchy file format [40]) output file that is lastly used to display avatar motion with a specialized Unity-based player software.

3.1 Dataset

The base of our work builds a collection of 2,632 annotated sign videos (average length ≈ 500 frames) from three independent signers captured with a resolution of 60Hz and their corresponding synchronized motion capture files. All videos only show the upper body of the respective signers. Within the available video collection, 1,432 files constitute continuous expressions of a female signer with a length of 3 to 14 lexical items each. These data captures represent either full sentence expressions or short phrases taken from a daily life domain in Japanese SL and contain common grammatical variations such as non-manual expressions or the use of classifier predicates. Further details about the dataset can be found in Reference [3]. The remaining data captures constitute 1,200 shorter expressions with a length of one to three lexical items in Japanese SL from the medical domain collected under identical conditions and capture protocol. Six hundred of these

expressions were signed by a second female signer and 600 expressions by a male signer. In total, approximately 310 minutes of signing video material were available for system development. The synchronized motion capture data contains high-resolution information of all relevant body joints without occlusion, whereas the placements of markers were the same for all three signers.

Two types of data were used for system training and testing in the following. These are 2D joint positions obtained from the collected sign videos utilizing the OpenPose framework p_{jk} and 3D absolute joint positions P_{jk}^* obtained from the synchronized motion capture data. The absolute positions were expressed in the global coordinate system and computed utilizing the relative rotations and translations of all skeletal joints as defined by the motion trajectories of the motion capture markers.

3.2 2D Data Pre-Processing

The 2D output data obtained utilizing the OpenPose network can be noisy and contain errors. Especially the hands, which move extremely quickly during a signed conversation, are often subject to missing samples caused by motion blur, occlusions between hands and lost tracking of joints. Therefore, all 2D pose estimates are first pre-processed before further use. Presuming that the legs do not move considerably during signing, the 2D position p_{jk} of knee and ankle joints that were not visible in the input video are roughly determined based on the length of the signer's hip ℓ_h and length of the signer's torso ℓ_t in image space. For example, the knee position $p_{k,r}$ and the ankle position $p_{a,r}$ of the right leg are approximated from the right hip position $p_{h,r}$ onward as

$$p_{k,r} = p_{h,r} + \begin{bmatrix} \ell_h * \lambda_{k_x} \\ \ell_t * \lambda_{k_y} \end{bmatrix} \text{ and } p_{a,r} = p_{k,r} + \begin{bmatrix} \ell_h * \lambda_{a_x} \\ \ell_t * \lambda_{a_y} \end{bmatrix}. \quad (3)$$

Here, λ_{k_x} and λ_{a_x} constitute scalars for the average human length ratio between thigh and shank bone along a video image's x-axis and λ_{k_y} and λ_{a_y} scalars for the average human length ratio between thigh and shank bone along a video image's y-axis. Next, missing data samples of the extracted joint trajectories are filled utilizing a simple 1D linear interpolation. Last, we design all computations in such a way that new video input can be processed successfully, even if a signer is unknown or camera properties differ. For this, we apply the following 2D pose rescaling strategy.

We know that a two-dimensional representation of a segment connecting two joints in three-dimensional space is of maximal length when it is parallel to the horizontal camera plane. In this way, let us assume that for a well-balanced and continuous movement, the frame f^* with maximal length for a line segment $\vec{v}_{j_1,j_2}$ of two points p_{j_1} and p_{j_2} in 2D space indicates the moment when $\vec{v}_{j_1,j_2}$ is perfectly parallel to the horizontal camera plane. This provides us with the individual segment lengths of a signer, with the length ℓ for any $\vec{v}_{j_1,j_2}$ being defined by

$$\ell = \max \sqrt{p_{j_1}^2 + p_{j_2}^2}. \quad (4)$$

Although ℓ constitutes a rough approximation of the actual segment length of a video's signer, experimentation shows that we can use the predicted lengths well to unify the 2D joint positions of the different signers. This makes it easier to relate the input of a subsequent pose estimation network to the training and target data. Utilization of the final system for the creation of new 3D skeleton data then imposes the existence of one sign language content video as the minimal requirement.

4 PROPOSED POSE ESTIMATION

The essential part of the pose estimation are three separate RNNs for inference of 3D joint positions of stick figure, face and finger in a global coordinate system with the x-axis representing the depth axis, the y-axis representing the horizontal axis, and the z-axis representing the vertical axis

Table 2. Basic Specifications of the Three Employed RNNs

	Sub-Pose	# Joints	# Channels	Initialization
RNN ₁	2D upper body stick figure	12	36	0, Y, Z
RNN ₂	2D finger	42 (21×2)	126	0, Y, Z
RNN ₃	2D face	32	96	0, Y, Z

in the upward direction. RNNs are chosen due to their ability of learning and storing sequential information. This constitutes an essential property for the reproduction and estimation of trajectory curves and resembles the properties of previous Kalman Filter-based approaches.

Each network has a distinct input/output layer size based on its respective number of joint positions to be estimated (Figure 4). For every RNN, all joints are treated as three dimensional, whereas the input constitutes the 2D positions of stick figure, finger, and face joints, respectively, further extended by the depth axis with values set to 0 (Table 2). Here, the idea is that positional information along the missing depth dimension should be comparatively easy to infer for a signing movement, during which many joints only undergo smaller movements on average. This leads to the following layer sizes: 36 feature channels (12 upper body stick figure joints $\times$ 3) for RNN₁, 126 feature channels (21 finger joints per hand $\times$ 3) for RNN₂, and 96 feature channels (32 facial joints $\times$ 3) for RNN₃.

4.1 Network Architecture

We implemented all of our networks in Python using TensorFlow version 1.8.0 and CUDA version 8.0.61. Several combinations of RNN cell structures and types (Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) in uni- and bidirectional versions), layers (3, 4, and 5), batch size (16, 32, 64), and number of hidden cells (128, 256, 512) were tested. The following configurations gave the best results and were hence applied in the final system for all three of our RNNs: batch size, 16; number of layers, 3; cell type, LSTM with layer normalization [1]; number of hidden cells, 512. The best RNN cell structure was a stacked bidirectional RNN with ReLU activations [17] as shown in Figure 5. This might be since for any linguistic problem, the words at the end of the sentence can be just as important as the words at the beginning in terms of providing context. Applying the same logic to our problem, the signs in the past and/or future may likewise help dictate the depth dimension or stabilize the data.

4.2 Network Training

For all RNNs, the 3D absolute joint positions obtained from the motion capture data serves as ground-truth data during the network training process. Approximately 90% of all data files per signer were randomly chosen from the full data collection of 2D joint positions to be used as training data. This led to a total training data size of 2,400 files and a test data size of 232 files. Each model was trained on a Nvidia GeForce GTX 1080 Ti and varied in terms of duration, depending on the dataset. For RNN₁, training took approximately between 6 to 10 days, while RNN₂ and RNN₃ took approximately between 2 to 4 days to converge. We used two different loss functions for each model: Mean Absolute Error (MAE) and Mean Squared Error (MSE). However, all of our models performed better with MAE, thus all results discussed further will refer to MAE unless stated otherwise.

4.3 Network Output

The outputs of our RNNs are sequences of 3D movement in the form of absolute positional data in the global coordinate system. All three networks are able to stabilize the data, whilst predicting the

Fig. 5. In our network, separate bidirectional RNNs (BiRNNs) that fold out into the forward (F) and backward directions (B) are stacked one on top of the other. The output of the previous BiRNN is the input for the next.

depth dimension based only on horizontal and vertical information. Outputs from RNN_1 furthermore appear much more realistic and stable than the stick figure estimates previously obtained with the 3d-pose-baseline model (see Figure 3 in Section 2.1). No obvious unnatural deformations are present and the network is able to keep the stick figure in place. Putting the outputs of RNN_1 , RNN_2 and RNN_3 together, we obtain a point cloud of a high fidelity, humanoid figure signing in 3D that closely resembles its baseline (Figure 6).

5 AVATAR MOTION COMPUTATION

The estimated 3D pose output can be used as is for SLR settings or other purely data-based applications. However, to generate sign avatar animations, further data transformation is required. A virtual sign language avatar can easily be animated once the angular and positional displacements of all joints of its skeletal hierarchy are known. We utilize the inferred absolute joint positions to compute such displacements following the IK calculus introduced in Reference [12]. For this, we build a custom-made robotic model that represents a simplified version of the given avatar skeleton restricted to the biomechanic constraints of the human body (Figure 7). For example, both elbows are modeled as joints of 1 degree of freedom (DOF) and a range of rotation between 0 and 150 degrees. This prevents the avatar from making non-humanoid movements. Simultaneously, it reduces the amount of potential movements and hence facilitates computation of the joint angles. To depict fine-grained movements that are important for the natural impression of the sign avatar, we additionally include transitional links for all facial joints. For example, the lower jaw is modeled as a combination of one rotational link to simulate the opening of the mouth as well as three transitional links to simulate positional changes along the three joint axes. This makes the required IK calculus the largest ever tested within the underlying robotics framework. In concrete, our model

Fig. 6. Visual comparison between point cloud figures of inferred and actual motion capture data. Each row shows one frame of a test sequence sample under a front and side view. First and third columns: combined output of the RNNs. Second and fourth columns: absolute positions obtained from the motion capture data.

allows us to simultaneously compute 232 DOFs, of which 52 rotations refer to the estimated stick figure positions and 66 rotations refer to the estimated finger positions. The remaining 114 displacements (93 rotations and 21 translations) depict the inferred changes of facial expression.

The IK algorithm aligns all joints of the kinematic model to their designated corresponding 3D joint position and subsequently determines their most likely angular and transitional displacements. In the last step, the computed displacements are merged into a BVH file [44] for avatar animation according to the hierarchical structure of the avatar skeleton.

6 EXPERIMENTS

The application targeted in this work constitutes a very unique research problem. For this reason, no common dataset, nor a common baseline technology, is available that could be used for evaluation of the proposed system. Therefore, we determine the accuracy of our system output in terms of absolute framewise similarity between the generated data and the real motion capture data. Additionally, we assess the system usability with respect to its intended subsequent application.

6.1 Evaluation: Absolute Difference

Similar as for the stick figure errors discussed in Section 2.1, we determine the deviation between the estimated 3D pose data P_{jk}^e of all test sequences and the absolute joint positions P_{jk}^* computed from the motion capture data to obtain a simple measure of the generated sequence quality. Following Equation (2), we consider the L_1 distance between its feature vector s representing the rescaled output of any of the three RNNs with size i and the feature vector g containing the corresponding

Fig. 7. The custom-made kinematic model used to compute joint angular displacements of the sign avatar's skeleton. Left: neutral standing pose. Right: IK computation. The white balls designate the reference points obtained with the pose estimation and are used to align the joints accordingly.

ground-truth positions with size i . For RNN_1 , we obtain an average score of $d_{R1} = 0.808$ cm absolute stick figure pose distances. The variance is $\sigma_{R1} = 1.229$ cm and the right wrist's x (depth) dimension is the most unstable with an average distance of 2.113 cm. Comparing the average differences to the one obtained with the normalized and rescaled output of the 3d-pose-baseline model, all values show a considerable improvement in accuracy. For example, the average distance of the most unstable wrist position is already $\approx 41\%$ smaller than the average distance d_{3dBL} computed over all joints. The total average distance d_{R1} is $\approx 78\%$ smaller than d_{3dBL} , whereas this comparison should even be biased in favor of the 3d-pose-baseline model: d_{3dBL} was composed of all stick figure joints, including the static legs, while d_{R1} only takes into account the relevant, moving upper body joints, which are likely to be of larger error than all joints that do not undergo positional changes. For the finger and face joints, average and maximal distances are smaller than for RNN_1 . This should also be necessary to correctly represent all fine-grained sub-parts of the overall movement. For RNN_2 , we obtain an average sequence distance $d_{R2} = 0.682$ cm with $\sigma_{R2} = 1.173$ cm variance. The right index fingertip's x dimension had the largest difference at 1.509 cm. For RNN_3 , we obtain an average sequence distance $d_{R3} = 0.695$ cm with $\sigma_{R3} = 0.725$ cm variance. Here, the center jaw's x dimension had the largest difference at 1.199 cm.

Qualitative evaluation between both generated curves shows that the network is generally able to well reproduce the target curves (Figure 8). Smaller difficulties mainly arise with respect to the location of the dynamically moving wrist, and in particular during movements that undergo highly

Fig. 8. Comparison of inferences (orange curves) to their target (blue curves), where the testing data index, joint and axis are noted above. The rows correspond to the outputs of RNN_1 , RNN_2 , and RNN_3 , respectively. For each RNN, we examine the curves of the joints with the largest absolute distance along the x -axis. The left column corresponds to inferences that have the smallest absolute difference, while the right column corresponds to the ones that have the largest. For better visual display, the scaling of all curve plots varies in dependence of their overall joint displacement.

frequent and sharp turns as well as the opening and closing of the mouth and eyes. Whether and how these might influence and change the semantic meaning of a signed expression should be examined under a more quantitative perspective in the application context.

6.2 Evaluation: SLR Scenario

We investigate the quality of our synthesized 3D joint trajectories in the SLR context. Here, the idea is that, if a neural network trained on motion capture data is able to reliably classify synthesized data, we can draw more meaningful conclusions about the semantic similarity of both data. This especially holds true for neural networks trained in a pure-data-driven way: Specifics of independent lexical items (sign words, respectively classes) learned from implicit features within the training data should then also be present within the generated data, while its registered deviations do not lead to confusion with other words.

For this reason, we train a simple Convolutional Neural Network (CNN) that classifies separate word segments occurring within a signed sequence. The word segments of a training set $\mathcal{T}$ comprise absolute 3D pose trajectories obtained with the motion capture system and are cut on the

Table 3. Accuracy Metrics of the CNN Trained on Real Motion Capture Data for the Two Evaluation Sets $\mathcal{V}1$ and $\mathcal{V}2$

	Accuracy	Precision	Recall	F1
$\mathcal{V}1$	0.8649	0.83	0.86	0.84
$\mathcal{V}2$	0.8268	0.80	0.83	0.80

basis of manual annotations. To evaluate our system performance with unknown video input, we only use the word segments cut from all previous training sequences within $\mathcal{T}$. Word segment annotations for all previous test sequences are used to build two separate evaluation datasets, $\mathcal{V}1$ constituting the cuts of the motion capture data test sequences, and $\mathcal{V}2$ constituting the cuts of the inferred pose trajectories for the test sequences. In this way, we are able to directly compare the classifier's performance differences with respect to unseen real and generated input data.

Since temporal annotations were only available for the collection of the 1,432 longer sign sentences [3], we only use the respective data and obtain 9,425 word segments of 288 different lexical items (classes) for evaluation. Our CNN architecture consists of two sets of one convolutional layer with ReLU activation (kernel size $[5 \times 5]$ and filter numbers [16] and [32], respectively) and a pooling layer (kernel size $[2 \times 2]$) followed by two fully connected layers (filter numbers [1024] and [288], respectively) with a final softmax activation. We utilize a batch size of 300 and apply dropout of 0.5 to the fully connected layers during training. Since our objective is to assess the quality of our generated data, we do not undertake any further fine-tuning of the network's hyper parameters. All of the position data is normalized by rescaling the corresponding joint's kinematic dependencies with respect to the maximal and minimal translation of every body segment. To build our imagelike data structure, we simply feed all normalized joint positions (stick figure, finger and face joints) into a feature matrix following the order of the network outputs. No further data transformations, such as the Joint Angular Displacement Maps proposed by Kumar et al. [33] or any of their further developments, are performed and varying data lengths are handled by zero-padding all word segments to 164 frames. Experimentation has shown that data transformations raise the overall classifier accuracy to more than 95% in dependence on the amount and type of data used. Since the overall difference between $\mathcal{V}1$ and $\mathcal{V}2$ remains similar for all tested variations, only the simplest classifier variation should be discussed here to keep the used data as close to a future real-time scenario as possible.

The classifier converges to an accuracy of 99.33% under the training set $\mathcal{T}$ in less than 1500 epochs. We compare the trained classifier's performance for the two test datasets with respect to the general wordwise accuracy and the precision, recall and F1 score averaged over all words (Table 3). $\mathcal{V}1$ has an accuracy of 86.49% and a F1 score of 0.84, $\mathcal{V}2$ reaches an accuracy of 82.68% and a F1 score of 0.80. This constitutes an absolute difference of 3.81% accuracy, or a relative drop in accuracy by 4.4%. Similarly, the overall absolute difference among precision, recall, and F1 score constitutes $\approx 3\%$ on average.

A closer look at the specific output predictions reveals that the majority of all words could be classified correctly. Differences with respect to the estimated labels of all word segments within $\mathcal{V}1$ and $\mathcal{V}2$ can hardly be detected visually, e.g. when comparing their respective confusion matrices (Figure 9). Most misclassifications occur for classes that are very rare and occur less than 5 times within $\mathcal{T}$. Classes that are morphologically very similar to other classes, however, can be classified without difficulty as long as they occur sufficiently frequent (minimally 5–10 times) within $\mathcal{T}$. Examples of such similar classes are lexical items with identical base motion shape, but subtle modifications induced by linguistic features, such as questions and negations (movement of

Fig. 9. Confusion Matrices for $\mathcal{V}1$ (left) and $\mathcal{V}2$ (right).

the head) or intonation (changes within facial expression). This indicates that the most relevant characteristics of a word are all represented within the generated data and can be well maintained from the video input throughout the data generation process.

6.3 Evaluation: SLS Scenario

The synthesis of high quality sign avatars largely relies on optical motion capturing, and the provision of 3D skeleton data from an automated video processing pipeline could considerably accelerate the development of SLS research. Evaluation of the comprehensibility of a sign avatar's movement is furthermore the most meaningful way to assess the quality and semantic correctness of the synthesized data.

To start, we visually compare the newly synthesized sequences to their corresponding motion captured test sequences. For this, we first determine the joint angles as described in Section 5. Next, we utilize the resulting BVH files containing all relevant joint displacements to replay the sequences with our given sign avatar model. Overall, we discover many successful avatar animations, but also animations with subtle errors. Figure 10 illustrates some of these success and anomaly cases. As already assumed from the previous evaluations, our system generalizes reasonably well and is able to synthesize smooth avatar movements that maintain the flow of the complete signed expressions as well as the main morphological shape of single linguistic items. This includes the correct representation of complicated interactions between multiple body parts, such as clapping, hitting the palm of the non-dominant hand with the dominant hand or touching different parts of the body. Previous analysis of the 3D joint trajectories revealed largest difficulties for the reproduction of highly frequent wrist movements. These present themselves in a subtle softening of the overall avatar movement and a lack of sharp and sudden turns. In some cases, deviations of the wrist position leads to a partial penetration of fingers when both wrists or fingers are very close to each other. Last, facial movements that do not match the motion capture data can be discovered. These were not noticeable in a 3D point cloud (see Figure 6). One reason for this is that an observer's perception is very sensitive to abnormalities in human expression [19, 50]. A fully natural movement of the avatar's facial expression might be obtained with system parameter fine-tuning. In particular, slight adaptations in the rescaling values applied to the output trajectories of RNN_3 , or adaptations of the translational ranges of the IK model, could already sufficiently soften the final avatar animation.

Fig. 10. Qualitative comparison between avatar animations using the synthesized data (first and third rows) and the real motion capture data (second and fourth rows) under a front and side perspective. The basic shape of the signs is well preserved. Deviations from the original are most noticeable within the avatar's face.

As a last step, we perform a user-based evaluation of our avatar animations to examine how the discovered differences might influence the overall perception and comprehensibility of the avatar. Twenty-five native speakers of Japanese SL with an age range of 13–68 years (mean: 39.08 years $\pm$ 16.33 years) volunteered to participate in the evaluation (9 male, 16 female). All participants self-identify as deaf or hard-of hearing, possess general familiarity with virtual characters and are experienced in use of technology such as handheld devices. No participant had ever seen animations of a sign avatar before. Following evaluation standards of previous investigations [23, 27–29], we assess the user rating with respect to four quality aspects for comprehensibility, naturalness and personal affinity on a 5-point Likert-scale:

- Q1) Was the signing easy to understand? (1 = confusing, 5 = clear)
- Q2) Was the signing natural? (1 = moves like robot, 5 = moves like a person)
- Q3) Was the signer realistic? (1 = not, 5 = very realistic)
- Q4) Did you like the signer? (1 = hate it, 5 = love it)

Every participant was asked to evaluate 120 video animations. These animations comprised a randomized selection of 60 pairs of motion capture and synthesized data BVH files from the set of test sequences. The videos were evaluated consecutively with the help of a specialized test program, whereas the order of appearance of all videos was further randomized and unknown to the participants. Averaging the scores of all motion capture animations and all synthesized animation separately, we can then draw a comparison that is independent of the content of the signed expression.

User evaluations show an obvious drop in the overall score per participant for all four evaluated questions (Figure 11). In the context of the previous analysis, this decrease is larger than expected and seems a heavy punishment of the rather small deviations found in the motion trajectories. However, our results are inline with previous examinations on sign avatar acceptance. A large-scaled study by Kipp et al. [29] for example showed a considerable quality gap between purely hand-made animations and automated avatars. On a 5-point scale ranging from –2 to +2,

Fig. 11. Visualization of the average user ratings per quality aspects Q1–Q4 and their standard deviation for the motion captured and synthesized sign animations.

Table 4. Minimal, Maximal, and Mean Evaluation Scores with Their Standard Deviation of the Motion Capture and Synthetic Data Driven Avatar Animations over All Evaluated Sentences and Quality Aspects Q1–Q4

	Q1		Q2		Q3		Q4	
	MoCap	Synth.	MoCap	Synth.	MoCap	Synth.	MoCap	Synth.
min	1	1	1	1	1	1	1	1
max	5	5	5	5	5	5	5	5
mean	4.195	2.824	4.109	2.664	3.988	2.48	3.803	2.327
std	0.381	0.425	0.339	0.409	0.343	0.367	0.390	0.424

hand-made animations achieved average ratings between ≈ 0.6 to ≈ 1.3 for various quality expects of the signed animation, while synthesized avatar animations achieved average ratings between ≈ -0.7 to ≈ -0.4 . Fitting these values to our Likert scale, namely ≈ 3.6 to ≈ 4.3 for the hand-made animations and ≈ 2.3 to ≈ 2.6 for the synthetic avatars, both assessments show the same tendencies. An older study by Huenerfauth [21] that explores different timing models for sign animation generation similarly shows relatively low rating scores for all types of synthesized avatars. Under a 10-point Likert scale (1 to 10), best performing models achieve a score of ≈ 7.5 with respect to comprehensibility and ≈ 6.1 with respect to naturalness of the movement. Worst performing models achieve a score of ≈ 3.3 for comprehensibility and ≈ 4.1 for naturalness. This would correspond to ≈ 3.75 and ≈ 3.05 , respectively ≈ 1.65 and ≈ 2.05 , under a 5-point Likert scale.

In both evaluations of Kipp et al. and Huenerfauth, naturalness of the avatar is rated lower than other quality aspects of the avatar. We cannot observe the same effect for our study and obtain similar rating differences of ≈ 1.4 points for all quality aspects (Table 4). As considered within the overall drop of user affinity, the newly synthesized data therefore appears to convey the fundamental characteristics of the signed expressions equally well under all relevant aspects. Variance of the ratings for the synthesized data is slightly higher than the variance for the motion capture data. A more detailed comparison of the rating distributions shows the users' clear tendency toward higher (respectively better) scores for the motion capture data videos and toward lower (respectively worse) scores for the generated data videos (Figure 12), whereas both video types achieve the same minimal and maximal rating for all quality aspects. This rating variance suggests that the assessment of the generated sign animations is subject to overall differences in personal perception. Quality differences within the generated data sequences, as the ones shown in Figure 8, might affect the user perception. As stated above, human perception is very sensitive to anomalies in the facial expression of a humanlike agent or robot. For sign avatars, the same might also apply to the

Fig. 12. Visualization of the users' rating selections over all evaluated sentences for motion capture and synthetic data driven avatar animations by a diverging stacked error bar with the center rating score of 3 used as neutral response.

full appearance of a signed expression: even the high-resolution motion capture data, respectively its display by a high-quality avatar model, does not reach close to a perfect (5-point) score. This observation is supported by another study of Kipp et al. [28], which has shown that users clearly prefer the transmission of information via real sign videos over a transmission of information via avatar animations. Results of our user assessment should therefore be treated in the context of such prevalent and specific type of uncanny valley [43].

7 DISCUSSION

The synthesized data appears to be of large similarity to the motion captured data with respect to both absolute framewise positional differences and intrinsic semantic meaning as learned by a data-driven neural network. This indicates that the proposed pipeline is able to provide meaningful data for subsequent use in sign translation research. However, it is difficult to generalize our assumptions due to the lack of a baseline dataset. We analyzed the accuracy and usability of the proposed system in the context of a non-public data collection and the data split was performed in an arbitrary fashion that balanced the amount of each signer's video content in the test and training set. Given the long time of network convergence, no cross-validation could be performed. As a consequence, the choice of the used data as well as its randomization might have a minor influence on the obtained results. However, such potential perturbations should not have a large impact on the general observations discussed in the following.

Analyzing the pipeline under a common loss function of sequence similarity shows that the positional difference between generated and estimated trajectories is small throughout all joint positions. The 3D stick figure—the only set of joints that can be compared to existing methods—shows no deformations and the estimated pose of the signer is much more natural and stable than the pose estimated by established methods. This also makes its framewise deviations considerably smaller than the ones of a well-known baseline method. Sequence dissimilarity is largest for the wrist joints, which also undergo the largest positional changes during a signed expression. However, wrist movement potentially also allows for larger differences in its trajectories: deviations within a range of few centimeters might hardly be noted by the observer, nor influence the meaning of a sign once the main morphological shape is maintained. To address the observed difficulty of estimating sudden, high frequent changes of direction, it might help to try different training losses that put more emphasis on extreme values in the target curves and punish softened estimates in respective areas. Variations should be explored in the following to make the estimation more robust.

For the additional SLR scenario evaluation, classification results of $\mathcal{V}2$ suggest that the generated data represents the characteristics of individual motion classes in a similar way as the motion capture data. Taking into account the number of word classes, as well as the unbalanced word distribution within the training data, we believe that a drop of $\approx 3\%$ accuracy does not indicate a relevant loss of performance strength of the trained classifier. In particular, it is important to note that the classifier maintained its ability to distinguish between small subtleties in a motion that represent linguistic features such as questions, negations or adjective inflection. They contain important information for the correct translation of a signed sentence, but are difficult to train within continuous SLR system. This is because those features commonly do not induce any changes of the general morphological shape of the motion class (as represented in the positional changes of the upper body stick figure joints), and only vary with respect to subtle changes of the head movement or changes of the facial expressions. The possibility to reliably estimate such fine-grained motion properties substantiate the importance of our proposed pipeline as compared to existing approaches that only provide 3D stick figure estimates of a (signing) person. To draw a final conclusion about the importance of the synthesized data, it would additionally be good to next utilize the pipeline as data generation tool for continuous SLR. Data augmentation can act as a regularizer and prevent overfitting [10, 48], which is especially important for neural networks trained on a restricted number of data. Since to date, any existing SL dataset is small (as compared to speech and language datasets), additional high-dimensional motion data synthesized from accessible video content could considerably enhance robustness and accuracy of translation systems in the future. A previous related examination on the dataset used in this article has shown that significant improvements of the classifier accuracy could be registered once the augmented training corpus contained more than 8,000 continuous sentences in total [4]. Consequently, a respective study would require a collection of some thousand annotated motion captured sentence expressions (to serve as basic training data) plus several thousand annotated sign videos of the same SL and content domain (to serve as augmentation data obtained from the proposed pipeline). Currently, such data is not available, so that a likewise evaluation could not be performed in this article.

Evaluation of the additional SLS scenario shows that we are able to animate a sign avatar with the computed angular and translational joint displacements in a similar way than with the actual motion capture data. The avatar correctly conveys the content of a signed expression, while following a smooth motion flow that includes natural facial expression and accurate finger movements. To achieve the best possible 3D skeleton data for the avatar synthesis of new content input, it is furthermore possible to fine-tune independent sub-parts of the final system. Evaluation by native speakers punishes the final animation of the avatar driven by synthesized data. However, comparing the scores to literature, our avatar achieves similar ratings as previous sign avatar generations, indicating that the semantic content of the signing is largely conveyed correctly. One cause for the reduction in user scores might be the intermediate IK mapping step. This procedure constitutes the only known (non machine learning based) method to determine joint displacements of a skeleton structure necessary for animation of the virtual avatar and can hardly be avoided. With its largest ever tested amount of DOFs, the simplified model furthermore already imposes challenging numerical computations, so that extension of the model to the actual avatar skeleton structure is not expected to improve accuracy of the underlying IK framework's output. Last, it should be emphasized that all synthesized data animations were only put into relation to a quasi-perfect baseline animation obtained from the motion capture data. Comparison to avatar animations generated by conventional sign synthesis strategies could change the user assessment in favor of our avatar model. In particular, we believe that naturalness of the movement and personal affinity could largely benefit from such a comparison, since our model

offers a higher degree of model- and render-based realism than existing avatars. In this context, we also want to emphasize that all participants had no previous experience with sign avatars, and that their rating might change with previous exposure and familiarity of sign animations.

8 CONCLUSION

In this work, we presented a novel pipeline for the generation of 3D skeleton data from sign videos for subsequent use as data augmentation tool in sign translation scenarios. In particular, we deployed three separate stacked bidirectional Recurrent Neural Networks to estimate the 3D positions of a signer's body, finger and facial joints in global coordinate space from previously extracted 2D positions. In the future, this data can be used as additional training data for continuous sign language recognition scenarios, which conventionally demand a huge amount of training data to handle all high-dimensional aspects of a signed conversation. We next utilized a kinematic robot model to compute angular and translational joint displacements of a sign avatar's skeleton via Inverse Kinematics, which could be subsequently utilized to drive the animation of a sign avatar. This could facilitate the collection of data resources for the generation of sign avatars, and hence improve sign avatar quality and user acceptance in the future. Consequently, our video processing pipeline comprises two major output data: the estimated absolute 3D joint positions of a sign video and motion capture data files of a sign video built from the determined joint displacements.

We evaluated our system with respect to sequence similarity and in two meaningful application contexts. Results show that we are able to learn motion sequences of high similarity to the real motion capture data. As far as we know, to date no other system architecture has been reported that estimates similarly detailed, accurate and reliable 3D joint trajectories of a detailed human skeleton beyond basic stick figure body shapes. Moreover, the generated data appears to contain and preserve all relevant morphological changes that are necessary for utilization in both sign recognition and a sign synthesis scenarios. This constitutes a valuable contribution to the sign language research society, which is suffering under the chronic lack of highly detailed data that could be acquired in an easy and cheap way. The 3D pose estimation step of our pipeline utilizes data normalized in a standardized procedure and is based on an existing framework that allows 2D tracking of stick figure, finger and face of multiple persons. Hence, any arbitrary video of sign language input, including multi-signer videos, should be successfully processed in the future.

As a next step, we plan to further assess the quality of our synthesized avatar animations utilizing the results of the performed user study. Besides, we intend to apply our pipeline under its primary purpose, the generation of data from sign videos as found on video sharing sites or as taken from public television broadcasts, to optimize the final system settings. Given meaningful outputs, we then ultimately hope to be able to utilize our system for the future development of a more accurate and robust sign translation system.

APPENDIX

A JOINT TRAJECTORIES

Figure 8 gives a visual comparison of the target and estimated joint trajectories for the body, face and finger joint of largest displacement along the x -axis. Since the x -axis is estimated from 2D data, and hence a zero state along the depth axis, it is subject to the largest differences within the axes for all 3 RNNs. To illustrate every system aspect, the following Figures 13 and Figure 14 show the

Fig. 13. Comparison of inferences (orange curves) to their target (blue curves) for the y trajectory outputs of RNN_1 , RNN_2 , and RNN_3 for the joints with the largest absolute distance. Following the arrangement given in Figure 8, the left column corresponds to inferences that have the smallest absolute difference, while the right column corresponds to the ones that have the largest absolute difference.

corresponding results for the y - and z -axis, respectively. Differences between target and inference curves are smaller than for the x -axis and follow the same trajectory pattern. Here, it should be noted that the scaling of all curve plots varies in dependence of their overall joint displacement for better visual display.

Fig. 14. Comparison of inferences (orange curves) to their target (blue curves) for the z trajectory outputs of RNN_1 , RNN_2 , and RNN_3 for the joints with the largest absolute distance. Following the arrangement given in Figure 8, the left column corresponds to inferences that have the smallest absolute difference, while the right column corresponds to the ones that have the largest absolute difference.

REFERENCES

- [1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
- [2] Agathe Balayn, Heike Brock, and Kazuhiro Nakadai. 2018. Data-driven development of virtual sign language communication agents. In *Proceedings of the 2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'18)*. IEEE, 370–377.
- [3] Heike Brock and Kazuhiro Nakadai. 2018. Deep JSLC: A multimodal corpus collection for data-driven generation of Japanese sign language expressions. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC'18)*. European Language Resources Association (ELRA).
- [4] Heike Brock, Juliette Rengot, and Kazuhiro Nakadai. 2018. Augmenting sparse corpora for enhanced sign language recognition and generation. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018) and the 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community*. European Language Resources Association (ELRA).
- [5] Patrick Buehler, Mark Everingham, Daniel P. Huttenlocher, and Andrew Zisserman. 2011. Upper body detection and tracking in extended signing sequences. *Int. J. Comput. Vis.* 95, 2 (2011), 180.
- [6] Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, and Richard Bowden. 2017. Subunets: End-to-end hand shape and continuous sign language recognition. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV'17)*. IEEE, 3075–3084.
- [7] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. OpenPose: Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7291–7299.

- [8] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7291–7299.
- [9] James Charles, Tomas Pfister, Mark Everingham, and Andrew Zisserman. 2014. Automatic and efficient human pose estimation for sign language videos. *Int. J. Comput. Vis.* 110, 1 (2014), 70–90.
- [10] Dan Claudiu Cireşan, Ueli Meier, Luca Maria Gambardella, and Jürgen Schmidhuber. 2010. Deep, big, simple neural nets for handwritten digit recognition. *Neur. Comput.* 22, 12 (2010), 3207–3220.
- [11] Thomas Coogan, George Awad, Junwei Han, and Alistair Sutherland. 2006. Real time hand gesture recognition including hand segmentation and tracking. In *Proceedings of the International Symposium on Visual Computing*. Springer, 495–504.
- [12] Behzad Dariush, Michael Gienger, Bing Jian, Christian Goerick, and Kikuo Fujimura. 2008. Whole body humanoid control from human motion descriptors. In *Proceedings of the 2008 IEEE International Conference on Robotics and Automation*. IEEE, 2677–2684.
- [13] Sarah Ebling and John Glauert. 2016. Building a Swiss German Sign Language avatar with JASigning and evaluating it among the Deaf community. *Univ. Access Inf. Soc.* 15, 4 (2016), 577–587.
- [14] Ali Farhadi, David Forsyth, and Ryan White. 2007. Transfer learning in sign language. In *Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.
- [15] Jens Forster, Christoph Schmidt, Oscar Koller, Martin Bellgardt, and Hermann Ney. 2014. Extensions of the sign language recognition and translation corpus RWTH-PHOENIX-weather. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC'14)*. 1911–1916.
- [16] Srujana Gattupalli, Amir Ghaderi, and Vassilis Athitsos. 2016. Evaluation of deep learning based pose estimation for sign language recognition. In *Proceedings of the 9th ACM International Conference on Pervasive Technologies Related to Assistive Environments*. ACM, 12.
- [17] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 6645–6649.
- [18] Junwei Han, George Awad, and Alistair Sutherland. 2009. Automatic skin segmentation and tracking in sign language recognition. *IET Comput. Vis.* 3, 1 (Mar. 2009), 24–35. DOI : <https://doi.org/10.1049/iet-cvi:20080006>
- [19] Jessica Hodgins, Sophie Jörg, Carol O'Sullivan, Sang Il Park, and Moshe Mahler. 2010. The saliency of anomalies in animated human characters. *ACM Trans. Appl. Percept.* 7, 4, Article 22 (Jul. 2010), 14 pages.
- [20] Jie Huang, Wengang Zhou, Qilin Zhang, Houqiang Li, and Weiping Li. 2018. Video-based sign language recognition without temporal segmentation. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence (2018)*.
- [21] Matt Huenerfauth. 2008. Evaluation of a psycholinguistically motivated timing model for animations of American sign language. In *Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility*. ACM, 129–136.
- [22] Matt Huenerfauth. 2008. Generating American sign language animation: Overcoming misconceptions and technical challenges. *Univ. Access Inf. Soc.* 6, 4 (2008), 419–434.
- [23] Matt Huenerfauth, Liming Zhao, Erdan Gu, and Jan Allbeck. 2008. Evaluation of American sign language generation by native ASL signers. *ACM Trans. Access. Comput.* 1, 1, Article 3 (May 2008), 27 pages.
- [24] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2014. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 7 (2014), 1325–1339.
- [25] Aaron S. Jackson, Adrian Bulat, Vasileios Argyriou, and Georgios Tzimiropoulos. 2017. Large pose 3D face reconstruction from a single image via direct volumetric CNN regression. In *Proceedings of the IEEE International Conference on Computer Vision*. 1031–1039.
- [26] Hernisa Kacorri and Matt Huenerfauth. 2016. Continuous profile models in asl syntactic facial expression synthesis. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 2084–2093.
- [27] Hernisa Kacorri, Matt Huenerfauth, Sarah Ebling, Kasmira Patel, Kellie Menzies, and Mackenzie Willard. 2017. Regression analysis of demographic and technology-experience factors influencing acceptance of sign language animation. *ACM Trans. Access. Comput.* 10, 1 (2017), 3.
- [28] Michael Kipp, Alexis Heloir, and Quan Nguyen. 2011. Sign language avatars: Animation and comprehensibility. In *Proceedings of the International Workshop on Intelligent Virtual Agents*. Springer, 113–126.
- [29] Michael Kipp, Quan Nguyen, Alexis Heloir, and Silke Matthes. 2011. Assessing the deaf user perspective on sign language avatars. In *Proceedings of the 13th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS'11)*. ACM, New York, NY, 107–114.
- [30] Oscar Koller, Sepehr Zargaran, Hermann Ney, and Richard Bowden. 2018. Deep sign: Enabling robust statistical continuous sign language recognition via hybrid CNN-HMMs. *Int. J. Comput. Vis.* 126, 12 (2018), 1311–1325.

- [31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 1097–1105.
- [32] Jack B. Kuipers et al. 1999. *Quaternions and Rotation Sequences*. Vol. 66. Princeton University Press, Princeton, NJ.
- [33] E. Kiran Kumar, P. V. V. Kishore, A. S. C. S. Sastry, M. Teja Kiran Kumar, and D. Anil Kumar. 2018. Training CNNs for 3-D sign language recognition with color texture coded joint angular displacement maps. *IEEE Sign. Process. Lett.* 25, 5 (2018), 645–649.
- [34] François Lefebvre-Albaret, Sylvie Gibet, Ahmed Turki, Ludovic Hamon, and Rémi Brun. 2013. Overview of the Sign3D project high-fidelity 3D recording, indexing and editing of French sign language content. In *Proceedings of the 3rd International Symposium on Sign Language Translation and Avatar Technology (SLTAT'13)* 2013.
- [35] Sijin Li and Antoni B. Chan. 2014. 3d human pose estimation from monocular images with deep convolutional neural network. In *Proceedings of the Asian Conference on Computer Vision*. Springer, 332–347.
- [36] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuoling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. MediaPipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172* (2019).
- [37] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. 2017. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 2640–2649.
- [38] John McDonald, Rosalee Wolfe, Jerry Schnepf, Julie Hochgesang, Diana Gorman Jamrozik, Marie Stumbo, Larwan Berke, Melissa Bialek, and Farah Thomas. 2016. An automated technique for real-time production of lifelike animations of American Sign Language. *Univ. Access Inf. Soc.* 15, 4 (2016), 551–566.
- [39] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. 2017. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Trans. Graph.* 36, 4 (2017), 44.
- [40] Maddock Meredith, Steve Maddock, et al. 2001. Motion capture file formats explained. *Department of Computer Science, University of Sheffield* 211 (2001), 241–244.
- [41] Thomas B. Moeslund and Erik Granum. 2001. A survey of computer vision-based human motion capture. *Comput. Vis. Image Understand.* 81, 3 (2001), 231–268.
- [42] Caio D. D. Monteiro, Christy Maria Mathew, Ricardo Gutierrez-Osuna, and Frank Shipman. 2016. Detecting and identifying sign languages through visual features. In *Proceedings of the 2016 IEEE International Symposium on Multimedia (ISM'16)*. 287–290. DOI : <https://doi.org/10.1109/ISM.2016.0063>
- [43] Masahiro Mori. 1970. The uncanny valley. *Energy* 7, 4 (1970), 33–35.
- [44] The University of Wisconsin. 1999. Biovision BVH. Retrieved from <http://research.cs.wisc.edu/graphics/Courses/cs-838-1999/Jeff/BVH.html>.
- [45] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. 2017. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7025–7034.
- [46] Nikolaos Sarafianos, Bogdan Boteanu, Bogdan Ionescu, and Ioannis A. Kakadiaris. 2016. 3d human pose estimation: A review of the literature and analysis of covariates. *Comput. Vis. Image Understand.* 152 (2016), 1–20.
- [47] Frank M. Shipman, Ricardo Gutierrez-Osuna, and Caio D. D. Monteiro. 2014. Identifying sign language videos in video sharing sites. *ACM Trans. Access. Comput.* 5, 4, Article 9 (Mar. 2014), 14 pages. DOI : <https://doi.org/10.1145/2579698>
- [48] Patrice Y. Simard, David Steinkraus, John C. Platt, et al. 2003. Best practices for convolutional neural networks applied to visual document analysis. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR'03)*, Vol. 3.
- [49] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1145–1153.
- [50] William Steptoe, Anthony Steed, Aitor Rovira, and John Rae. 2010. Lie tracking: Social presence, truth and deception in Avatar-mediated telecommunication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'10)*. ACM, New York, NY, 1039–1048.
- [51] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*. 3104–3112.
- [52] Bugra Tekin, Pablo Márquez-Neila, Mathieu Salzmann, and Pascal Fua. 2017. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3941–3950.

Received July 2019; revised November 2019; accepted December 2019