Exploiting 3D Hand Pose Estimation in Deep Learning-Based Sign Language Recognition from RGB Videos

Maria Parelli[1], Katerina Papadimitriou[2], Gerasimos Potamianos[2], Georgios Pavlakos[3], and Petros Maragos[1]

[1] School of ECE, National Technical University of Athens, Greece
[2] ECE Department, University of Thessaly, Volos, Greece
[3] GRASP Laboratory, University of Pennsylvania, Philadelphia, USA

maryparelli@gmail.com, {aipapadimitriou, gpotamianos}@uth.gr, pavlakos@seas.upenn.edu, maragos@cs.ntua.gr

Abstract. In this paper, we investigate the benefit of 3D hand skeletal information to the task of sign language (SL) recognition from RGB videos, within a state-of-the-art, multiple-stream, deep-learning recognition system. As most SL datasets are available in traditional RGB-only video lacking depth information, we propose to infer 3D coordinates of the hand joints from RGB data via a powerful architecture that has been primarily introduced in the literature for the task of 3D human pose estimation. We then fuse these estimates with additional SL informative streams, namely 2D skeletal data, as well as convolutional neural network-based hand- and mouth-region representations, and employ an attention-based encoder-decoder for recognition. We evaluate our proposed approach on a corpus of isolated signs of Greek SL and a dataset of continuous finger-spelling in American SL, reporting significant gains by the inclusion of 3D hand pose information, while also outperforming the state-of-the-art on both databases. Further, we evaluate the 3D hand pose estimation technique as standalone.

Keywords: sign language recognition, 3D hand pose, 2D body skeleton, attention-based encoder-decoder, convolutional neural network

1 Introduction

Automatic sign language recognition (SLR) from video has been attracting significant interest lately, following recent deep learning advances in the fields of computer vision and human language technologies, as well as the collection of suitable large SL corpora [4, 5, 34]. However, despite much progress in the field, the problem remains challenging due to the complex nature and multitude of SL articulators (both manual and non-manual), as well as variability in inter-subject signing and in the quality of the available video data.

A significant portion of SLR approaches in the literature utilize hand and/or body skeletal information in their pipelines. Such can be obtained from special

Proceedings European Computer Vision Conference Workshops (ECCVW-2020) - Workshop on Sign Language Recognition, Translation & Production (SLRTP-2020), Glasgow, UK, Aug. 2020. https://slrtp.com/

data acquisition tools in conjunction with wearable markers or data gloves [26, 28], but at the expense of naturalness in the interaction, or provided directly by RGB-D cameras [12, 20, 40] that also yield a depth information data stream. Specifically, systems utilizing such cameras have been introduced [11, 25, 35], with most of them relying on hand-crafted feature descriptors extracted from the depth and/or skeleton streams. For example, work in [19] explores the incorporation of 3D skeleton data, leveraging the advancement of depth sensors for SLR. Further, a number of promising works [23, 30] are based on 2D body skeletal data inferred from OpenPose [39]. Such representations are often complemented with appearance or optical-flow based motion information [21, 24, 36].

Since hand-based articulation plays a crucial role in SL and there exist multiple signs with very similar skeletal motion patterns, it is vital to seek schemes that enrich hand motion and structure information. The problem of hand pose estimation in videos is a long-standing one and has given birth to many important applications. Following the emergence of RGB-D sensors, many efforts have been devoted to 3D hand pose estimation through depth sensor and RGB input. However, in the majority of SL corpora and in real-world settings, signs are not recorded with depth sensors and depth information is unavailable. Thus, since RGB cameras are more widely used than depth sensors, recent works focus mostly on 3D hand pose estimation from monocular RGB images. The work in [44] is the first to address the problem with the use of deep learning, adopting a three-stage pipeline that performs hand segmentation, 2D joint generation, and then 3D joint prediction. A similar approach is suggested in [31], where state-of-the-art deep-learning networks are used for 2D hand detection and 2D hand joint localization, and the results are fitted to a generative model formulated as a non-linear least-squares optimization problem. In addition, the work in [29] proposes a cycle-consistent generative adversarial network (CycleGAN) which transforms synthetic 3D annotated hand images into real looking ones, whose statistical distribution matches real-world hand images. The resulting data are trained via a convolutional neural network (CNN) regressor for 2D and 3D hand joint predictions, and the predictions are fitted to a kinematic skeleton model. Finally, one of the most recent advances is the work of [17], which proposes a hand-model regularized graph refinement network for 3D hand pose estimation from a monocular image. It employs an adversarial learning framework and estimations from a parametric hand model as a structure prior, which is then refined via residual graph convolution.

In this paper, we incorporate the depth dimension in the coordinates of the hand joints, in order to enrich model knowledge about the trajectory of hand movement by enabling its observation in 3D. Our motivation is that such enriched information, effectively capturing the relative position between hand joints in the 3D space, will translate to improved SLR performance. To this end, and as detailed in Section 2, we extract 3D hand skeletal information exclusively from RGB videos through a powerful architecture [27], originally proposed in the field of 3D human pose estimation. Specifically, after extracting 2D human skeleton data of the body, hands, and face via the OpenPose library [39], we

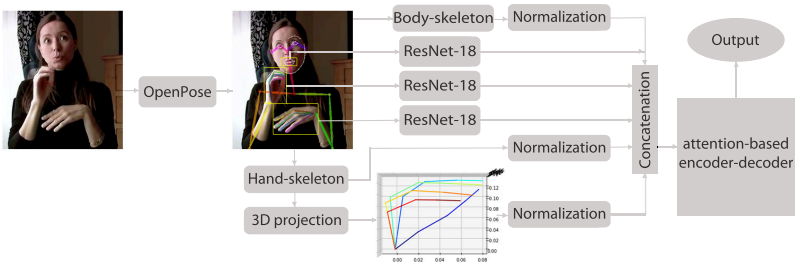


Fig. 1. Architecture of our proposed SLR system operating on RGB videos. Estimated 3D hand pose features are concatenated with additional SL informative feature streams and fed to an attention-based encoder-decoder for SLR.

project 2D hand-joint coordinates to the 3D space via a deep multi-layer neural network [27]. We then utilize an appropriately normalized representation of the 3D hand-joint estimates for SLR, in conjunction with state-of-the-art attention-based encoder-decoder architectures for sequence-to-sequence prediction. Further, we include more SL informative streams in the SLR system, in order to investigate the additional benefit of the 3D hand pose. Specifically, we consider normalized 2D skeletal features, as well as CNN-based representations via the ResNet-18 architecture [16] of the hands and mouth regions-of-interest (ROIs), segmented based on the 2D skeletal information, capturing manual articulation (handshape) and mouthing information, respectively. To our knowledge, this constitutes the first ever investigation of 3D hand pose information within a state-of-the-art, multiple-stream, deep learning-based SLR framework operating on traditional RGB video data.

We conduct SLR evaluations on two suitable multi-signer datasets: (i) a corpus of isolated signs of Greek SL (GSL) [2] and (ii) the ChicagoFSWild database [37], namely a corpus of continuous finger-spelling in the American SL (ASL). On both sets, inclusion of 3D hand pose information is able to benefit SLR on top of all other feature streams combined. Further, our results exceed the current state-of-the-art on both sets. In addition, we report experimental results of the 3D hand pose estimation technique on the Rendered HandPose [44] and the FreiHAND [45] datasets. Details are provided in Section 3.

2 The Sign Language Recognition System

We next overview our proposed SLR system, also depicted in Fig. 1, providing details of its feature extraction and sequence-to-sequence prediction modules, as well as its implementation details.

2.1 Feature extraction

2D human skeleton detection and features: The system initiates with the extraction of 2D human skeletal data employing the OpenPose human-joint de-

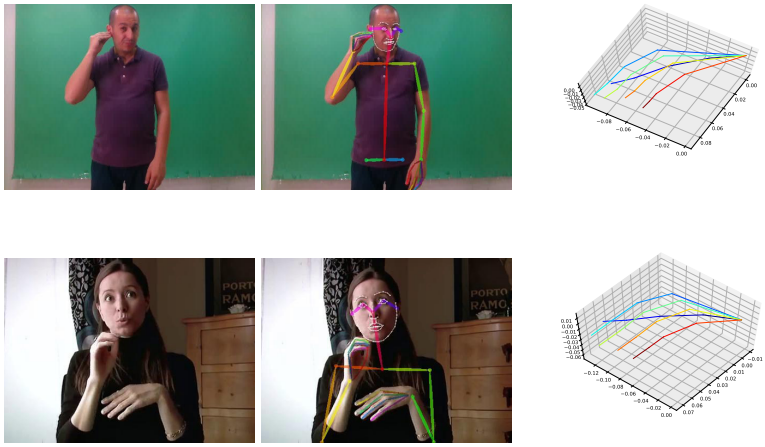


Fig. 2. Examples of extracted human skeleton via OpenPose [39] (middle column) and 3D hand skeleton representation through 3D projection architecture [27] (right column) on original data frames (left column). Images are from the GSL dataset (upper row) and the ChicagoFSWild corpus (lower row).

tector [39], which provides a descriptive motion and structural representation of the human body, employing deep convolutional pose models. Specifically, OpenPose renders in total 137 body skeleton joint descriptors, extracted in the form of image pixel coordinates, namely 25 body pose keypoints, 21 joints for each hand, and 70 face keypoints, as also depicted in the middle column of Fig. 2. Since only upper-body videos are employed in this work, we exploit 57 extracted image coordinates, excluding 10 body joints corresponding to the lower body, as well as the face keypoints. As a result, 114-dimensional (dim) feature vectors are extracted capturing the 2D coordinates of the upper-body skeleton (30-dim) and the two hands (84-dim in total). Note that, to incorporate translation and scale invariance, the estimated 2D human skeletal joints are subjected to normalization by transforming them to a local coordinate system, where the neck joint is assumed to be the origin, whereas further normalization is applied based on the distance between the left and right shoulder keypoints.

2D to 3D hand skeleton projection: Our approach extracts 3D hand-joint keypoints by "lifting" 2D joint locations to the 3D space. Our input is a series of 2D hand-joint keypoints, previously generated by the OpenPose framework, and our output is a series of points in the 3D space. We zero-center both 2D and 3D poses around the wrist joint, so as to ensure that our model learns translation-invariant representations. A noticeable source of error in 3D joint predictions is noise in the input 2D predictions. Since an increase in performance is noticed when smoothing is applied to the input, we use a median filter with radius one to remove noise spikes and eliminate instability in the predictions.

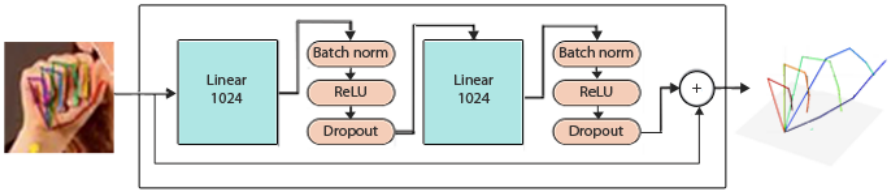


Fig. 3. The building blocks of the architecture that generates 3D hand skeleton joints from 2D hand skeletal data (figure modified from [27]).

Consequently, we implement a simple but powerful architecture, originally proposed in [27] for human pose 3D estimation, also depicted in Fig. 3. Our model is a deep multi-layer neural network with batch normalization, dropout, rectified linear units (ReLU), and residual connections. The latter improve generalization performance, while batch normalization and dropout improve model robustness to noisy 2D detections. Additionally, in order to further stabilize performance, a constraint on the weights of each layer is applied, so that the maximum norm is less than or equal to one. More precisely, the building block of the network is a linear layer followed by batch normalization, dropout, and ReLU activation. This block is repeated twice, and the two blocks share a residual connection. For this task we stack two outer residual blocks, and our model contains approximately 4 million trainable parameters. For network training, we use the Rendered HandPose Dataset [44], a large-scale 3D hand pose dataset based on synthetic hand models (see also Section 3.1).

The model yields 21 3D joints for each hand, thus 126-dim feature vectors are extracted. Note that, for translation and scale invariance, the wrist is assumed as the coordinate system origin, and the hand 3D keypoints are further subjected to normalization according to the distance between the shoulder and elbow keypoints of each hand.

Hand and mouth ROIs extraction and appearance features: Hands contain the most prominent SL information. Additional information also exists in mouthing patterns, being part of non-manual SL articulation. Thus, our system detects the ROIs of the mouth and each hand, exploiting the corresponding 2D human skeleton points returned by OpenPose. To generate appearance feature maps (one for the mouth and one for each hand), each ROI is resized to 224×224 pixels and fed to a ResNet-18 network [16] (using 3×3 convolutional kernels and downsampling with stride 2), pretrained on the ImageNet corpus [9]. This yields 512-dim features for each stream by taking the output of the network fully-connected layer.

Feature fusion: The extracted feature streams are then fused through simple vector concatenation. Thus, our SLR system employing all aforementioned data

streams will have 1776-dim features (114-dim for the 2D human skeleton, 126-dim for the 3D hand joints, and 512-dim for the ROIs of each of the mouth and two hands). Additional systems with fewer feature streams (hence lower dimensionalities) are also evaluated in Section 3.4. It should be noted that in case of missing streams due to OpenPose failures or occluded hands, the respective features are filled by zeros.

2.2 Sequence learning model

Regarding SLR from videos as a sequence-to-sequence prediction task, we address the SLR problem by a sequence learning approach based on an encoder-decoder module equipped with an attention mechanism. In its general form, the encoder is fed with the latent representations generated by a particular feature learner outputting a hidden states sequence, which is then processed by the decoder producing the predicted output. Further, the attentional models are based on the alignment between input and output accomplished by the likelihood of each portion of source sequence being related to the ongoing output.

Considering the above typical structure, a variety of attention based encoder-decoder schemes have been proposed, with most of them being mainly associated with recurrent neural networks (RNNs). The most dominant RNN encoder-decoder variants are long short-term memory networks (LSTMs) [18] and gated recurrent units (GRUs) [6]. Additionally, various architectures have been introduced relying on bi-directional RNNs [3, 42, 43]. Recently, the Transformer multi-head attention-based architecture [41] has been proposed that instead of involving CNNs or RNNs, it is complemented with position encoding and layer normalization. Moreover, in [32], a sequence-learning model using multi-step attention-based CNNs (enabling parallelization) is employed for finger-spelling recognition.

In this work, four sequence learning models are considered, namely: an attentional LSTM encoder-decoder [18], an attentional GRU encoder-decoder [6], an attentional CNN encoder-decoder [32], and a Transformer network [41]. Details of their implementation are provided next.

2.3 Implementation Details

All aforementioned deep-learning models are implemented in PyTorch [33], and their training carried out on a GPU. Specifically, the 3D hand skeleton generation network of Section 2.1 is trained for 150 epochs using the Adam optimizer [22], a batch size of 64, a starting learning rate of 0.001 and exponential decay. The weights of the linear layers are set using Kaiming He initialization [15].

For the sequence-learning models of Section 2.2, we employ a one-layer attentional LSTM encoder-decoder [18] with 128 hidden units and a one-layer GRU encoder-decoder [6] with hidden dimensionality equal to 256. Both RNNs are trained via the Adam optimizer [22] with an initial learning rate of 0.001 decayed by a factor of 0.3 and a dropout rate of 0.3. Beam search is applied during decoding with beam-width 5. The attentional CNN encoder-decoder model has 3

layers with kernel width 5 and 256 hidden units, and its training is based on the Adagrad optimizer [10] with an initial learning rate of 0.003, decreased by a factor of 1.0. Dropout of 0.1 and beam search of width 5 are employed. Finally, the Transformer is a 4-layer one with 8 heads for Transformer self-attention, 2048-dimension hidden Transformer feed-forward, and 512 hidden units. Its training is conducted via the Adam optimizer with an initial learning rate of 0.001 decreased by a factor of 2.0 and dropout 0.4. Parameter initialization is carried out by the Xavier process [14].

3 Experiments

Before proceeding to the SLR experiments that constitute the main focus of this paper, we briefly evaluate our 3D hand pose estimation approach.

3.1 3D hand pose corpora

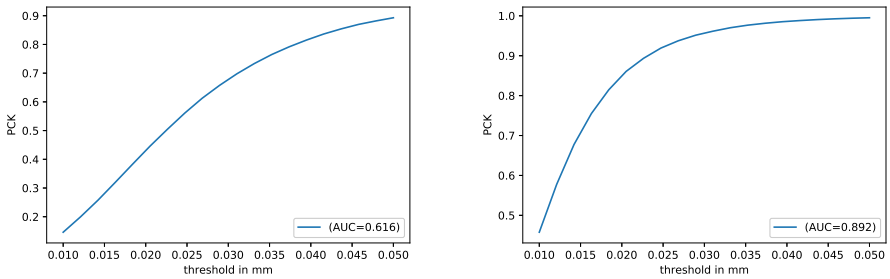
We conduct experiments on the 3D hand skeleton generation network performance using two corpora: the Rendered HandPose dataset (RHD) [44] and the FreiHand database (FHD) [45]. More details are provided next.

Rendered HandPose dataset: We use this corpus for network training. It constitutes a large-scale 3D hand pose dataset, based on synthetic hand models [44]. The dataset utilizes 3D human models with corresponding animations from Mixamo 2 [13], and the open-source software Blender 3 [7] is used for image rendering. It consists of 20 different characters performing 39 actions, and for each frame a different camera location is randomly selected. The dataset provides 41,258 images for training and 2,728 images for evaluation with a resolution of 320×320 pixels. Annotations of a 21 keypoint skeleton model of each hand are available, as well as segmentation masks. In our work, we take advantage of the hand keypoints with their coordinates in the image frame and their coordinates in the world frame.

FreiHAND Dataset: The dataset consists of real images and shows samples both with and without object interactions. It is captured with a multi-view setup and contains 33,000 samples. Hand poses are recorded from 32 subjects, and the set of actions include ASL signs, counting and moving fingers to their kinematic limits. 3D annotations for 21 hand keypoints are provided. For this work, we partition the data to 80% for training and 20% for testing.

3.2 3D hand pose estimation results

In Table 1 and Fig. 4, we evaluate the performance and generalization power of our model on the aforementioned datasets for various training / testing scenarios, reporting average median point error per keypoint of the predicted 3D pose,



(a) (b)

Fig. 4. Percentage of correct keypoints (PCK) over a certain threshold in mm, evaluated: (a) on RHD-test for model trained on RHD-train; and (b) on FHD-test for model trained on FHD-train.

when given the 2D ground truth pose, as well as the area under the curve (AUC) on the percentage of correct keypoints for different error thresholds. Specifically, in Table 1, among other results, in order to investigate the cross-dataset generalization of our network, we use the model trained on RHD-train and report AUC score and median error per joint on the FHD dataset, after alignment with the ground truth (Procrustes analysis). We also report percentage of correct keypoints (PCK) in Fig. 4, which returns the mean percentage of predicted joints below an Euclidean distance from the correct joint location.

The results show that our method demonstrates good performance on both datasets. The RHD set is characterized as challenging, due to the variations in viewpoints, and as a result we report higher 3D pose error. Since we are mostly interested in the generalization power of our model and its performance "in the wild", we find that our model manages to adapt effectively to unseen data and accurately captures the hand pose.

Table 1. Performance of the 3D hand pose estimation algorithm evaluated by two metrics for different training/testing scenarios.

Metrics ⟹		AUC	Median error
Training	Testing	score	per joint (mm)
RHD-train	RHD-train	0.729	18.1
	RHD-test	0.616	22.6
	FHD-test	0.771	16.2
FHD-train	FHD-test	0.900	11.0

3.3 SL Corpora

As already mentioned, the proposed SLR system is evaluated on two multi-signer SL corpora: (i) The isolated sign GSL dataset [2] and (ii) the continuous ASL fingerspelling ChicagoFSWild database [37]. More details are provided next.

GSL dataset: This consists of 15 (5×3) different dialogues, organized in sets of 5 individual tasks in 3 public services, performed by 7 different signers. The dialogues, between a deaf person and a single service employee, are pre-defined and are performed by each signer 5 consecutive times ($5 \times 7 \times 5 \times 3$). Signing is captured by an Intel RealSense D435 RGB-D camera at a rate of 30 Hz, providing simultaneously RGB and 24-bit depth data streams at the same spatial resolution of 648×480 pixels. Additionally, during recording, camera pose adjustments are made, thus offering a desirable variation in the videos. Corpus annotations by GSL linguistic experts are provided at both the signed sentence and signed word levels. The corpus signed vocabulary consists of 310 unique glosses (40,785 gloss instances) and 331 unique sentences (10,290 sentences), with 4.23 glosses per sentence on average. Here, an isolated sign recognition task is built concerning 306 unique words (numerals are discarded) that are expressed between 4 and 10 times by each signer in the dataset, yielding 12,897 clips. The dataset is trained under a multi-signer framework, with all experiments conducted through ten-fold cross-validation, where 80% of each fold is allocated to training, 10% to validation, and 10% to testing.

ChicagoFSWild database: This corpus includes ASL finger-spelling image frame sequences collected from online videos, providing a natural SL corpus in a real-world setting. The absence of unique signs for several words, such as names, foreign lexical items, and technical terms renders finger-spelling [32, 37, 38] a meaningful SL variant, basically expressed in a continuous letter signing unscrambling manner. The corpus was annotated through ELAN [1, 8] by students that have studied ASL. The data contain 7,304 ASL finger-spelling sequences with frame resolution of 640×360 expressed by 160 signers, leading to a 3,553 unique finger-spelled word vocabulary. Here, we employ a small-vocabulary subset concerning 103 unique finger-spelled words, involving 26 English letters with a sufficient number of occurrences among all signers (143 signers) between 10 and 130 times in the corpus. These yield 3,076 video snippets of words obtained by the ELAN annotation time-stamps of the words of interest. Training is conducted under a multi-signer setting, through ten-fold cross-validation with 80% of each fold used for training, 10% for validation, and 10% for testing. For comparison purpose, training is also conducted in a signer-independent (SI) setting, where the dataset is divided into training, validation, and testing sets without signers overlap among the partitions. Applying the same partition as in [37], the training partition corresponds to 5,455 samples, the validation 981 videos, and the testing set 868 clips.

Table 2. Word accuracy (%) on two SL datasets under a multi-signer experimental paradigm, employing various feature stream combinations in conjunction with the attentional multi-step CNN encoder-decoder sequence-learning model of Section 2.2.

Feature Streams					Datasets	
Hand CNNs (1024-dim)	Mouth CNN (512-dim)	2D-Hand Skeleton (84-dim)	2D-Body Skeleton (30-dim)	3D-Hand Skeleton (126-dim)	GSL	ChFSWild
✓					88.25	84.71
	✓				29.46	23.57
		✓			40.33	34.20
			✓		33.87	30.07
				✓	78.91	75.29
✓		✓			88.96	86.47
✓		✓	✓		93.17	90.81
✓			✓	✓	90.20	86.33
✓		✓	✓	✓	93.40	91.17
✓	✓				89.13	86.54
✓	✓		✓		89.81	87.65
✓	✓	✓	✓		93.41	91.01
✓	✓		✓	✓	91.23	87.10
			✓	✓	81.22	80.36
		✓	✓	✓	83.42	80.98
✓	✓	✓	✓	✓	**94.56**	**91.38**

3.4 SL recognition results

We first evaluate our SLR model for various feature streams using the attentional multi-step CNN encoder-decoder sequence-learning model, showcasing the power of 3D hand skeleton representations in the SLR task. Both datasets are evaluated in terms of word accuracy (%) in a multi-signer setting. As deduced from Table 2, the 3D hand skeleton seems to be a robust representation, achieving the highest accuracies on both datasets when added to 2D skeleton joints (body and hand skeletons) and hand and mouth articulator appearance feature representations, revealing the benefit of using multiple visual features streams that are complementary to each other. Incorporating 3D hand pose information boosts system performance on top of all other streams, obtaining 94.56% accuracy on the GSL dataset and 91.38% on ChicagoFSWild. It can also be viewed that its incorporation as additional hand information performs better when included with the 2D hand skeletal data. As demonstrated, the CNN-based articulator feature representations perform well, while the mouth region is mostly complementary in benefiting other feature streams. Finally, it can be observed that skeletal features yield lower accuracies when used alone than appearance feature streams, demonstrating the need for their combined used.

Next, in Table 3, we investigate the performance of the various sequence-learning techniques of Section 2.2, when employing all feature streams discussed

(1776-dim). Again, word accuracy is reported on the two datasets under the multi-signer experimental paradigm. As it can be observed, the best results are obtained by the attentional CNN encoder-decoder, revealing its superiority to the considered alternatives. This is primarily due to the good learning ability of CNNs. It can be readily seen that the worst results for the GSL dataset are obtained by the Transformer encoder-decoder module, while for the ChicagoFS-Wild database by the attentional GRU encoder-decoder.

We also evaluated the performance of the proposed system employing the attentional CNN encoder-decoder under a speaker-independent experimental paradigm in terms of letter accuracy (%) on the ChicagoFSWild dataset, improving over the best reported results of [38] from 45.1% to 47.93%. Additionally, our model outperforms previous reported approaches regarding the GSL dataset [2], yielding word accuracy improvements from 89.74% to 94.56%.

Further, in Table 4, a number of variations of the sequence-learning model (attentional CNN encoder-decoder) are considered, regarding the number of layers, kernel widths, and the beam width employed during decoding. Results demonstrate that deeper architectures enhance model performance.

Finally, it should be noted that our system was evaluated using alternative skeletal joints normalization schemes, namely instead of normalizing 3D hand skeletal data regarding the elbow-shoulder distance, we applied the shoulder-to-shoulder distance, achieving 1.23% less accuracy in the GSL dataset and 2.47% in the ChicagoFSWild dataset. Additionally, employing the Euclidean distance between joints, generating 57 instead of 114 2D skeletal features results in an accuracy decrease for both datasets (6.44% for GSL dataset and 8.91% for the ChicagoFSWild database).

4 Conclusion

In this paper we investigated the benefit of estimated 3D hand skeletal information to the task of SLR from RGB videos, within a state-of-the-art deep-learning recognition system, operating on multiple feature streams. We proposed to infer 3D hand pose from 2D skeletal information obtained from OpenPose, using a deep-learning architecture previously used for 3D human pose estimation. Our results on two multi-signer SL corpora demonstrated that 3D hand pose adds value on top of other feature streams, including 2D skeletal information and

Table 3. Word accuracy (%) on two SL datasets under a multi-signer experimental paradigm, using various encoder-decoder models with all feature streams concatenated.

Encoder-decoder	GSL corpus	ChFSWild corpus
Attentional LSTM	89.97	86.42
Attentional GRU	89.55	84.50
Attentional CNN	**94.56**	**91.38**
Transformer	88.21	85.63

Table 4. Comparative evaluation of model variations of the attentional multi-step CNN encoder-decoder sequence-learning model of Section 2.2 in terms of word accuracy (%) in a multi-signer setting, with L being the number of layers, KW the kernel widths, and BW the beam width.

Model details			Datasets	
L	KW	BW	GSL	ChFSWild
1	3	3	84.21	82.74
2	3	3	93.83	87.45
3	3	3	94.33	90.87
1	5	5	87.52	85.48
2	5	5	93.27	91.12
3	5	5	**94.56**	**91.38**

CNN-based representations of manual and non-manual articulators. Further, our results outperformed the previously reported state-of-the-art on the two SL corpora considered.

Acknowledgements

The research work was supported by the Hellenic Foundation for Research and Innovation (H.F.R.I.) under the "First Call for H.F.R.I. Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment grant" (Project Number: 2456).

References

1. ELAN (Version 5.8) [Computer software] (2019), Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. https://archive.mpi.nl/tla/elan
2. Adaloglou, N., Chatzis, T., Papastratis, I., Stergioulas, A., Papadopoulos, G., Zacharopoulou, V., Xydopoulos, G.J., Atzakas, K., Papazachariou, D., Daras, P.: A comprehensive study on sign language recognition methods. IEEE Transactions on Multimedia (2019)
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. CoRR **abs/1409.0473** (2014)
4. Camgoz, N.C., Hadfield, S., Koller, O., Ney, H., Bowden, R.: Neural sign language translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7784–7793 (2018)
5. Camgöz, N.C., Koller, O., Hadfield, S., Bowden, R.: Sign language transformers: Joint end-to-end sign language recognition and translation. CoRR **abs/2003.13830** (2020)
6. Cho, K., Merrienboer, B.V., Gülçehre, C., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 1724–1734 (2014)
7. Community, B.O.: Blender - a 3D modelling and rendering package. Blender Foundation, Stichting Blender Foundation, Amsterdam (2018), http://www.blender.org

8. Crasborn, O., Sloetjes, H.: Enhanced ELAN functionality for sign language corpora. In: Proceedings of the Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora. pp. 39–43 (2008)
9. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
10. Duchi, J., Hazan, E., Singer, Y.: Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research **12**, 2121–2159 (2011)
11. Escalera, S., Gonzàlez, J., Baró, X., Reyes, M., Guyon, I., Athitsos, V., Escalante, H., Sigal, L., Argyros, A., Sminchisescu, C., Bowden, R., Sclaroff, S.: Chalearn multi-modal gesture recognition 2013: Grand challenge and workshop summary. In: Proceedings of the ACM on International Conference on Multimodal Interaction. pp. 365–368 (2013)
12. Fossati, A., Gall, J., Grabner, H., Ren, X., Konolige, K. (eds.): Consumer Depth Cameras for Computer Vision - Research Topics and Applications. Springer (2012)
13. Fuse, M.: Mixamo: Quality 3D Character Animation In Minutes (2015), [online]: https://www.mixamo.com
14. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the International Conference on Artificial Intelligence and Statistics. pp. 249–256 (2010)
15. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1026–1034 (2015)
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
17. He, Y., Hu, W., Yang, S.F., Qu, X., Wan, P., Guo, Z.: 3D hand pose estimation in the wild via graph refinement under adversarial learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
18. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computing **9**, 1735–1780 (1997)
19. Hosain, A.A., Santhalingam, P.S., Pathak, P., Kosecka, J., Rangwala, H.: Sign language recognition analysis using multimodal data. In: Proceedings of the IEEE International Conference on Data Science and Advanced Analytics. pp. 203–210 (2019)
20. Hu, Y., Zhao, H.F., Wang, Z.G.: Sign language fingerspelling recognition using depth information and deep belief networks. International Journal of Pattern Recognition and Artificial Intelligence **32**(06) (2018)
21. Kartika, D.R., Sigit, R., Setiawardhana, S.: Sign language interpreter hand using optical-flow. In: Proceedings of the International Seminar on Application for Technology of Information and Communication. pp. 197–201 (2016)
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014)
23. Ko, S., Son, J., Jung, H.: Sign language recognition with recurrent neural network using human keypoint detection. In: Proceedings of the Conference on Research in Adaptive and Convergent Systems. pp. 326–328 (2018)
24. Konstantinidis, D., Dimitropoulos, K., Daras, P.: A deep learning approach for analyzing video and skeletal features in sign language recognition. In: Proceedings

of the IEEE International Conference on Imaging Systems and Techniques. pp. 1–6 (2018)
25. Kurakin, A., Zhang, Z., Liu, Z.: A real time system for dynamic hand gesture recognition with a depth sensor. In: Proceedings of the European Signal Processing Conference. pp. 1975–1979 (2012)
26. Lee, B.G., Lee, S.M.: Smart wearable hand device for sign language interpretation system with sensors fusion. IEEE Sensors Journal **18**(3), 1224–1232 (2018)
27. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3D human pose estimation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2659–2668 (2017)
28. Mittal, A., Kumar, P., Roy, P.P., Balasubramanian, R., Chaudhuri, B.B.: A modified LSTM model for continuous sign language recognition using Leap Motion. IEEE Sensors Journal **19**(16), 7056–7063 (2019)
29. Mueller, F., Bernard, F., Sotnychenko, O., Mehta, D., Sridhar, S., Casas, D., Theobalt, C.: GANerated hands for real-time 3D hand tracking from monocular RGB. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 49–59 (2018)
30. Nugraha, F., Djamal, E.C.: Video recognition of American sign language using two-stream convolution neural networks. In: Proceedings of the International Conference on Electrical Engineering and Informatics. pp. 400–405 (2019)
31. Panteleris, P., Oikonomidis, I., Argyros, A.A.: Using a single RGB frame for real time 3D hand pose estimation in the wild. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision. pp. 436–445 (2018)
32. Papadimitriou, K., Potamianos, G.: End-to-end convolutional sequence learning for ASL fingerspelling recognition. In: Proceedings of the Annual Conference of the International Speech Communication Association. pp. 2315–2319 (2019)
33. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. In: Proceedings of the NIPS-W (2017)
34. Pitsikalis, V., Theodorakis, S., Vogler, C., Maragos, P.: Advances in phonetics-based sub-unit modeling for transcription alignment and sign language recognition. In: Proceedings of the IEEE Computer Vision & Pattern Recognition Workshops. pp. 1–6 (2011)
35. Ren, Z., Yuan, J., Zhang, Z.: Robust hand gesture recognition based on finger-earth mover's distance with a commodity depth camera. In: Proceedings of the ACM Multimedia Conference and Co-Located Workshops. pp. 1093–1096 (2011)
36. Roussos, A., Theodorakis, S., Pitsikalis, V., Maragos, P.: Dynamic affine-invariant shape-appearance handshape features and classification in sign language videos. Journal of Machine Learning Research **14**, 1627–1663 (2013)
37. Shi, B., Rio, A.M.D., Keane, J., Brentari, D., Shakhnarovich, G., Livescu, K.: Fingerspelling recognition in the wild with iterative visual attention. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5399–5408 (2019)
38. Shi, B., Rio, A.M.D., Keane, J., Michaux, J., Brentari, D., Shakhnarovich, G., Livescu, K.: American sign language fingerspelling recognition in the wild. Proceedings of the IEEE Spoken Language Technology Workshop pp. 145–152 (2018)
39. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4645–4653 (2017)
40. Tashev, I.: Kinect development kit: A toolkit for gesture- and speech-based human-machine interaction [best of the web]. IEEE Signal Processing Magazine **30**(5), 129–131 (2013)

41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems (NeurIPS) 30. pp. 5998–6008 (2017)
42. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., Dean, J.: Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR **abs/1609.08144** (2016)
43. Zhou, J., Cao, Y., Wang, X., Li, P., Xu, W.: Deep recurrent models with fast-forward connections for neural machine translation. CoRR **abs/1606.04199** (2016)
44. Zimmermann, C., Brox, T.: Learning to estimate 3D hand pose from single rgb images. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4913–4921 (2017)
45. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 813–822 (2019)