WW-CSL: A New Dataset for Word-Based Wearable Chinese Sign Language Detection

Fan Xu, Kai Liu, Yifeng Yang, and Keyu Yan*
Jiangxi Normal University
{xufan, 202241600074}@jxnu.edu.cn, yyf317162188@outlook.com, yankeyu0328@163.com

Abstract

Sign language is an effective non-verbal communication mode for the hearing-impaired people. Since the video-based sign language detection models have high requirements for enough lighting and clear background, current wearing glove-based sign language models are robust for poor light and occlusion situations. In this paper, we annotate a new dataset of Word-based Wearable Chinese Sign Languag (WW-CSL) gestures. Specifically, we propose a three-form (e.g., sequential sensor data, gesture video, and gesture text) scheme to represent dynamic CSL gestures. Guided by the scheme, a total of 3,000 samples were collected, corresponding to 100 word-based CSL gestures. Furthermore, we present a transformer-based baseline model to fuse 2 inertial measurement unites (IMUs) and 10 flex sensors for the wearable CSL detection. In order to integrate the advantage of video-based and wearable glove-based CSL gestures, we also propose a transformer-based Multi-Modal CSL Detection (MM-CSLD) framework which adeptly integrates the local sequential sensor data derived from wearable-based CSL gestures with the global, fine-grained skeleton representations captured from video-based CSL gestures simultaneously.

Keywords: Chinese sign language, wearable glove, flex sensor, gesture video

1. Introduction

Sign language can express semantic meaning through the gestures of fingers, hands, and facial expressions. Therefore, sign language is an effective body language to exchange information and express rich semantic messages for the hearing-impaired people. Sign language can be widely used in many fields such as health care sector, sign langue teaching, and culture communication (Jin et al., 2017). To address the communication barriers faced by hearing-impaired individuals and mitigate their social isolation, the development of automatic sign language translation and interpretation systems holds significant promise.

Over the years, numerous models have been developed to detect sign languages across various linguistic regions, including American sign language (Dreuw et al., 2008; Lee et al., 2020), Chinese sign language (Huang et al., 2018), German sign language (Koller et al., 2015), Japanese sign language (Sun et al., 2013), Arabic sign language (Aliyu et al., 2017), Italian sign language (Escalera et al., 2014), and Korean sign language (Yang et al., 2020).

The sign language detection models typically fall into two primary categories: video-based and wearable glove-based. Generally, video-based models face limitations associated with visual angles and susceptibility to environmental factors. To address these challenges, researchers have turned their attention to the development of smart wearable sign language interpretation platforms using machine learning techniques. For instance, Lee et al. (2020) introduced a sensor fusion-based American Sign Language (ASL) interpretation framework. In addition, despite these advancements, no prior work has effectively integrated the strengths of both video-based and wearable glove-based sign language gesture recognition methods simultaneously.

Recognizing the paucity of multi-modal Chinese Sign Language (CSL) resources, including both video-based and wearable glove-based datasets and detection models, we embarked on the development of a comprehensive, novel resource. We present WW-CSL, a Word-based corpus for Wearable Chinese Sign Language, thoughtfully curated to address this research gap. To amass this corpus, we design a wearable data glove equipped with 2 Inertial Measurement Units (IMUs) and 10 flex sensors. WW-CSL includes a total of 3,000 samples, each corresponding to 100 distinct word-based CSL gestures. Moreover, we introduce a transformer-based baseline model that fuses two IMUs and ten flex sensors for wearable CSL detection. Our approach marries the advantages of video-based and wearable glove-based CSL gesture recognition by presenting a novel multi-modal CSL detection (MM-CSLD) framework. This framework adeptly integrates the local sequential sensor data derived from wearable-based CSL gestures with the global, fine-grained skeleton representations captured from video-based CSL gestures.

This paper makes two significant contributions.

(1) We introduce an innovative wearable glove equipped with two IMUs and ten flex sensors, establishing a robust foundation for the development of computational models for CSL detection.

* Corresponding author.

LREC-COLING 2024, pages 17718–17724
20-25 May, 2024.

(2) We present a robust multimodal CSL detection baseline that integrates the sequential sensor data from local wearable-based CSL gestures with the comprehensive, fine-grained skeleton representations extracted from video-based CSL gestures, offering a unified and comprehensive approach to CSL detection.

2. Related Work

In this section, we describe the representative sign language corpus and its corresponding computational models.

2.1. Sign Language Corpus

In the past decade, several sign language corpus (e.g., video-based and wearable glove-based sign language gestures) has been released. There are many video-based sign language corpus, e.g., German sing language (Forster et al., 2012; Agris et al., 2008), American sign language (Neidle et al., 2012; Pugeault and Bowden, 2011), Chinese sign language (Chai et al., 2014; Huang et al., 2018), Greek sign language (Efthimiou and Fotinea, 2007), Polish sign language (Oszust and Wysocki, 2013), Argentinian sign language (Ronchetti et al., 2016), Arabic sign language (Shohieb et al., 2015).

At the same time, there are also some wearable glove-based sign language datasets, e.g., Chinese sign language (Wang et al., 2012), American sign language (Lee and Lee, 2018; Mummadi et al., 2018). However, the size of these wearable glove-based corpora is small. For example, there are only 7 predefined hand gestures in the wearable CSL gestures (Wang et al., 2012), e.g., hand being vertically lifted upwards, hand waving from left to right, etc. Similarily, there are only 24 and 27 predefined ASL gestures in Mummadi et al., 2018 and Lee and Lee, 2018, respectively.

2.2. Sign Language Models

We introduce video-based and weareble glove-based models for the sign language detection as following.

Video-based Models: Yin et al. (2015) proposed an SVM-based sign language classification model to integrate the concatenated sparse coding of frame fragments. Huong et al. (2015) preposed a principal components analysis-based method to recognize 25 Vietnamese gestures. Tharwat et al. (2015) proposed an Arabic sign language recognition model which adopts scale-invariant feature transformation to extract the effective hand feature points. Gupta et al. (2016) presented a KNN-based Indian sign language gesture recognition model, obtaining accuracy of 90% for the 26 gestures. Tornay et al. (2020) proposed a Kullback-Leibler divergence-driven HMM-based multilingual sign language recognition model. Recently, many studies focused on adopting deep learning-based technology to conduct continuous sign language recognition. For example, Cui et al. (2019) introduced a sign language recognition framework based on Bi-LSTM, and integrated optical flow feature and RGB-D data simultaneously. Niu and Mak (2020) designed a transformer-based approach to model the temporal relationship among different frames of gesture video.

Weareble Glove-based Models: Lee et al. (2020) proposed a one-handed sensor fusion of motion-based ASL interpretation framework based on a recurrent neural network to recognize 27 word-based ASL gestures. Lee et al. (2021) extracted 30 feature points from one-hand leap motion sensor to conduct ASL recognition, obtaining accuracy of 91.82%. Kurtoglu et al. (2021) adopted RF (Radio Frequency) sensors to obtain time-frequency, range-Droppler, and range-angle for sequential trigger sign classification, obtaining accuracy of 92% for 15 ASL words and 3 gross motor activities. Lee and Lee (2018) presented an SVM-based 26 ASL signs recognition through a data-glove consists of 1 IMU and 5 flex sensors. Mummadi et al. (2018) introduced a random forest-based 24 static ASL letters recognition through a data-glove consists of 5 IMU sensors.

3. Annotation Scheme

3.1. Werable Glove Design

Figure 1 illustrates our werable glove which adopts 2 inertial measurement unites (IMUs) and 10 flex sensors. The IMU is fixed to the back of wrist, and the flex sensors are attached to all fingers. 10 flex sensors are attached to the backs of 10 fingers, from the root to the tip, with the aim of collecting bending data of the metacarpophalangeal joints. When the metacarpophalangeal joint bends, the interphalangeal joint also bends accordingly, which greatly enhances the flexibility and durability of our gloves. The components of our wearable gove include MPU-6050, Raspberry Pi Zero 2W, and flex sensors. The IMU sensor releases outputs of acceleration from the accelerometer, angular rate from the gyroscope, and the flex sensors deliver flex value for each finger.

3.2. Three-form Format Data

Huang et al. (2018) released a large-scale video-based CSL gestures which include 500 popular Chinese words and 102 Chinese sentences. We selecte the most common 100 Chinese words from their video-based CSL gestures to generate our werable glove-based sequential sensor data. The

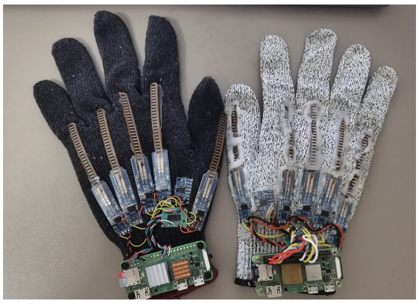

Figure 1: Our wearable glove.

100-word of CLS gestures include "situation", "objective", "clue", etc as shown in Table 1. We have three-form format data, e.g., sequential sensor data, gesture video, and gesture text. We collect 30 times of data for each CSL gesture using our werable glove, obtaing 3,000 samples.

Sequential Sensor Data: This sequential sensor data is a matrix consists of 200 lines and 22 colums for each CSL gesture which includes the acceleration in X-axis direction, acceleration in Y-axis direction, acceleration in Z-axis direction, magnetic field in X-axis direction, force angle, magnetic angle in the Y-axis direction, magnetic angle in the Z-axis direction, and the curvature of five fingers (from the thumb finger to thumb finger).

Gesture Video: Since our annotator wears data gloves, it is not feasible to collect the video format of our annotator when performing sign language gestures. To retain the same size of sequential sensor data, we also extract the initial 30 videos from Huang et al. (2018).

Gesture Text: The gesture text consists of the words related to a specific CSL gesture.

4. Werable Chinese Sign Language Corpus

In this section, we address the key issues of the WW-CLS annotation, including annotator training, quality assurance, and annotation instance.

4.1. Annotator Training

The annotator team consists of a Ph.D. in Chinese linguistics as the supervisor (senior annotator) and 1 graduate student who has studied sign language for more than 3 years as annotator. The annotation is done in three phases. In the first phase, the annotator spends 2 weeks on learning the principles of scheme. In the second phase, the annotator spends 2 months on annotating the 100 CSL gestures. In the final phase, the supervisor spends 2 weeks carefully proofread all 100 CSL gestures.

situation	part	objective
result	clue	gain
effect	degree	symbolize
popularize	develop	launch
occur	establish	solid
empty	tight	loose
inflate	children	youth
deaf	grand mother	father-in-law
neighbor	own	me
you	he	friend
male	female	mister
eye	ear	nose
mouth	heart	people
the people	model	nanny
security staff	writer	painter
student	kind	job
status	society	tradition
mop	knife	native place
carpet	ear pendants	necklace
bracelet	ring	soap
toothbrush	toothpaste	towel
watch	glasses	mirror
basin	cup	ruler
scissors	thread	switch
button	box	lock
key	umbrella	sauce
cigarette	moon	sun
steamed stuffed bun	wonton	cake
a baked pancake	Wotou	soup
snack	tin	preserved fruit
meat	egg	wine
tea	rain	bed
desk	chair	drawer
quilt		

Table 1: 100 CSL gesture words.

4.2. Quality Assurance

In order to generate diverse data, the annotator collects 30 times of data for each CSL gesture with fast/medium/slow speeds when performing these gestures. We calculate the average cosine similarity between the sequential sensor data (vector). The range of similarity value is from 0.50 to 0.90, and the average similarity value is 0.77. The wide range of similarity value guarantees the corpus's diversity. The more annotators, the higher the cost is.

4.3. An Annotation Instance

Table 2 presents a XML-style sample of our word-level CSL gestures. The <content> section demonstrates the specific contents including gestures con-

Table 2: A sample from our WW-CSL.

```
<?xml version="1.0" encoding="utf-8" ?>
<Content>situation</Content>
<Sensor>
7.2,-4.9,7.9, ... , 125.5
......
7.4,-3.7,5.1, ..., -11.8
</Sensor>
<Video>./GestureVideo0/0-29.mp4</Video>
</Gesture>
```

tent (i.e., words). The <sensor> section represents the corresponding sequential sensor data. The <video> section denotes the corresponding video for the CSL gestures.

5. Proposed Model

Figure 2 illustrates the framework of our transformer-based multimodal CSL detection (MM-CSLD). We adopt 1-layer encoder and 1-layer decoder of the transformer model. The encoder consists of a multi-head self-attention layer and a feedforward layer. The decoder has a query as the initial input, the output after passing through the multi-head projection module serves as the query of another multi-head attention module, and the output of the encoder serves as the key and value. Similar to the encoder, the last two modules of the decoder are the multi-head attention layer and the feedforward layer. Finally, the decoder passes through a linear layer with an output dimension of 100 to obtain the final Chinese sign language label. We select four skeleton points (e.g., 'HANDLET', 'HANDRIGHT', 'ELBOWLEFT', and 'ELBOWRIGHT') as inputs for the skeleton data from the gesture video.

The input of the proposed model is a fusion data F_i which fuses the sampled sensor data and skeleton data from gesture video as shown in Equation 1.

$$F_i = Concat(Flatten(S_i), Flatten(W_i)) \tag{1}$$

where $W = \{W_1, W_2, ..., W_t\}$ indicates the wearable-based sequential sensor data, $S = \{S_1, S_2, ..., S_t\}$ denotes the video-based skleton data, t embodies the time, $Concat$ represents concatation operation, and $Flatten$ refers to flattener operation which can flatten a matrix to a vector.

We fed the fusion data $F = \{F_1, F_2, ..., F_t\}$ into the position encoding and the vector encoded by the position into the encoder. We calculate the position encoding using Sin and Cosin functions as shown in Equations 2 and 3.

$$PE_{pos,2i} = sin\frac{pos}{10000^{\frac{2i}{d_{model}}}} \tag{2}$$

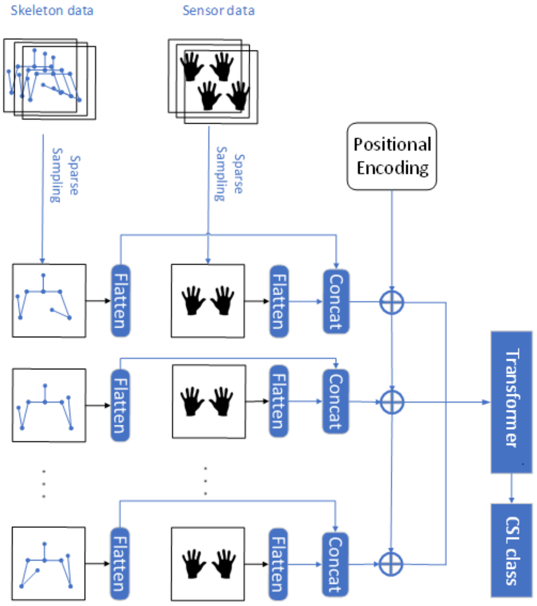

Figure 2: Our proposed MM-CLSD framework.

$$PE_{pos,2i+1} = cosin\frac{pos}{10000^{\frac{2i}{d_{model}}}} \tag{3}$$

where pos denotes current position, i represents the dimension index, and d_{model} indicates the size of dimension.

Data Preprocessing: For the sequential sensor data, we remove invalid data from all collected sensor data, perform sparse sampling on the sequences of unequal lengths, divide the sequence into an average of 16 parts, and randomly extract one time step of data from each data. Finally, the 16 time step data obtained is fed into our model as sensor data. For the skeleton data from a gesture video, we perform the similar process, obtaining equal time step of skeleton data.

6. Preliminary Experiments

6.1. Baselines

For the wearable-based CSL detection, we select GRU, LSTM, CNN, and Dense as baseline systems. We adopt 1-layer GRU and LSTM followed by 2-layer fully connection layer in the baseline modules. We adopt 1-layer CNN and 2-layer Dense followed by 1-layer fully connection layer in the baseline systems. For the video-based CSL detection, we adopt the CNN-based Conv3D (Tran et al., 2015) and the GRU-based FO-DMT-Net (Zhang et al., 2020) as our baseline modules.

6.2. Experimental Settings

Evaluation Metric: We adopt four popular evaluation metrics, i.e., accuracy, precision, recall, and F1-Score, to investigate the performance of our proposed framework and other comparing approaches.

Parameter Settings: The number of heads is set to 2 in the multi-head self-attention layer. The dimension of the feedforward layer is set to 1024, the activation function is RELU, and the dropout value is set to 0. The dimension size of the final linear layer is set to 100. The initial learning rate is set to e^{-4}, the batch size is set to 32, and training epochs is set to 300. The hidden size is set to 512 for the three baselines (e.g., Dense, LSTM, and GRU). The kernel size is set to 3 for the CNN baseline, and activation function is RELU. The input size of wearable-based sensor data is set to 22. The input size of video-based skeleton data is set to 12, and sample duration is set to 16. We adopt Adam optimizer to optimize our cross entropy loss function. Since the size of current corpus is limited, we follow previous works to omit validation set. We randomly select 80% for training and 20% for testing. Train 300 epochs, save the model once after each epoch, and all models converge after 300 epochs. Then, search for the best parameter model from the tensorboard utility and find the best performing model on the test set as the result of this model.

6.3. Experimental Results

Table 2 presents the performance comparision results. We can conclude that our proposed model MM-CLSD outperforms other baselines in terms of accuracy, precision, recall, and F1-Score measures. Our MM-CLSD successfully integrates the sequential sensor data from the local wearable-based and the global fine-grained skeleton representation for the video-based CSL gestures, respectively. Based on the findings in Table 3, we can derive the following insights:

(1) Our transformer-based model performs the best on the wearable-based sensor data. Intuitively, the sequential models (e.g, transformer, GRU, LSTM) perform better than non-sequential models (e.g., CNN, Dense) because sign language is a kind of time series.

(2) In fact, the Conv3D is a kind of CNN-based model, and the FO-DMT-Net is a kind of GRU-based model. They cannot capture the key sequential skleton data of sign language. In addition, the total number of skeleton of FO-DMT-Net is 25, resulting much noises in sign language detection.

(3) Since the dimension size of the input data equals to 34, only 2 or 17 heads can be adopted in the self-attention. The performance of using 17 heads is worse than that with 2 heads. The potential reason is that the large number of 17 heads cannot extract discriminative learning features to classify gesture type.

Table 3: Performance of CSL detection.

Model	Accu.	Prec.	Reca.	F1
Werable-based sensor data				
GRU	0.9183	0.9262	0.9183	0.9222
LSTM	0.9150	0.9279	0.9150	0.9214
CNN	0.8667	0.8862	0.8667	0.8763
Dense	0.8783	0.8994	0.8783	0.8887
MM-CSLD w/o video data	**0.9350**	**0.9449**	**0.9350**	**0.9399**
Video-based data				
GRU	0.5267	0.5615	0.5267	0.5435
LSTM	0.5933	0.6225	0.5933	0.6075
CNN	0.5733	0.5826	0.5733	0.5779
Dense	0.5283	0.5555	0.5283	0.5416
Conv3D	0.4450	0.4816	0.4450	0.4626
FO-DMT-Net	0.4810	0.6268	0.4810	0.5443
MM-CSLD w/o sensor data	**0.7033**	**0.7393**	**0.7033**	**0.7209**
Combined multimodal data				
GRU	0.9083	0.9262	0.9083	0.9172
LSTM	0.9317	0.9363	0.9317	0.9340
CNN	0.9133	0.9255	0.9133	0.9194
Dense	0.9083	0.9443	0.9083	0.9260
FO-DMT-Net	0.5267	0.6453	0.5183	0.5749
MM-CSLD(2 heads)	**0.9733**	**0.9779**	**0.9733**	**0.9756**
MM-CSLD(17 heads)	0.9650	0.9713	0.9650	0.9641

7. Conclusions and Future Work

We annotate a word-based wearable Chinese sign language dataset which adopts three-form format scheme by using our wearable glove consists of IMU and flex sensors. Meanwhile, we present a transformer-based multi-modal CSL detection framework on the proposed WW-CSL through integrating the local features from the wearable glove and the global features from the video gesture. Our future endeavors will focus on sentence-based CSL data collection and continuous CLS detection.

Acknowledgments

The authors would like to thank anonymous reviewers for their insightful comments on this paper. This research was supported by the National Natural Science Foundation of China under Grant 62162031 and 62066020, the Natural Science Fund project in Jiangxi province under Grant 20224ACB202010, and Funding of the "Jiangxi Province High level and Skilled Leading Talent Training Project".

Limitations

Our experiments were conducted on specific dataset (Chinese sign language) and may not fully represent the characteristics of other sign language detection scenarios or platforms. Generalizability to different datasets and languages needs to be further explored.

References

U.V. Agris, J. Zieren, U. Canzler, B. Bauer, and K.F. Kraiss. 2008. Recent developments in visual sign language recognition. *Universal Access in the Information Society*, 6(4):323–362.

S.O. Aliyu, M.A. Mohandes, and M. Deriche. 2017. Dual lmcs fusion for recognition of isolated arabic sign language words. In *In Proceedings of the International Multi-Conference on Systems, Signals & Devices (SSD'17)*, pages 611–614, Marrakech, Morocco.

X.J. Chai, H.J. Wang, and X.L. Chen. 2014. The devisign large vocabulary of chinese sign language database and baseline evaluations. In *Technical Report VIPL-TR-14-SLR-001*.

R.P. Cui, H. Liu, and C.S. Zhang. 2019. A deep neural framework for continuous sign language recognition by iterative training. *IEEE Transactions on Multimedia*, 21(7):1880–1891.

P. Dreuw, C. Neidle, V. Athitsos, S. Sclaroff, and H. Ney. 2008. Benchmark databases for video-based automatic sign language recognition. In *In Proceedings of the Conference on Language Resources and Evaluation (LRECâĂŸ08)*, pages 1–6, Marrakech, Morocco.

E. Efthimiou and S.E. Fotinea. 2007. Gslc: creation and annotation of a greek sign language corpus for hci. In *In Proceedings of the 4th International Conference on Universal Access in Human-Computer Interaction (UAHCI'07)*, pages 657–666, Beijing, China.

S. Escalera, X. Baró, J. Gonzàlez, M.A. Bautista, M. Madadi, M. Reyes, V.P. López, H.J. Escalante, J. Shotton, and I. Guyon. 2014. Chalearn looking at people challenge 2014: dataset and results. In *In Proceedings of the European Conference on Computer Vision (ECCV'14)*, pages 459–473, Zurich, Switzerland.

J. Forster, C. Schmidt, T. Hoyoux, O. Koller, U. Zelle, J. Piater, and H. Ney. 2012. Rwth-phoenix-weather: a large vocabulary sign language recognition and translation corpus. In *In Proceedings of the 8th International Conference on Language Resources and Evaluation (LRECâĂŹ12)*, pages 3785–3789, Online.

B. Gupta, P. Shukla, and A. Mittal. 2016. K-nearest correlated neighbor classification for indian sign language gesture recognition using feature fusion. In *In Proceedings of the International Conference on Computer Communication And Informatics (ICCCI'16)*, pages 1–5, Coimbatore, India.

J. Huang, W.G. Zhou, Q.L. Zhang, H.Q. Li, and W.P. Li. 2018. Video-based sign language recognition without temporal segmentation. In *In Proceedings of the Thirty-Second AAAI Conferenceon Artificial Intelligence (AAAI'18)*, pages 2257–2264, New Orleans, Louisiana, USA.

T.N.T. Huong, T.V. Huu, T.L. Xuan, and S.V. Van. 2015. Static hand gesture recognition for vietnamese sign language (vsl) using principle components analysis. In *In Proceedings of the International Conference on Communications, Management and Telecommunications (ComManTel'15)*, pages 138–141, DaNang, Vietnam.

C.M. Jin, O. Zaid, and J.M. Hisham. 2017. A review of hand gesture and sign language recognition techniques. *International Journal of Machine Learning and Cybernetics*, 10:131–153.

O. Koller, J. Forster, and H. Ney. 2015. Continuous sign language recognition: towards large vocabulary statistical recognition systems handling multiple signers. *Computer Vision and Image Understanding*, 141:108–125.

E. Kurtoglu, A.C. Gurbuz, E.A. Malaia, D. Griffin, C. Crawford, and S.Z. Gurbuz. 2021. Asl trigger recognition in mixed activity/signing sequences for rf sensor-based user interfaces. *arXiv preprint: arXiv:2111.05480v3*.

B. G. Lee, T.W. Chong, and W.Y. Chung. 2020. Sensor fusion of motion-based sign language interpretation with deep learning. *Sensors*, 20(6256):1–17.

B.G. Lee and S.M. Lee. 2018. Smart wearable hand device for sign language interpretation system with sensors fusion. *IEEE Sensors Journal*, 18:1224–1232.

C.K.M. Lee, K.K.H. Ng, C.H. Chen, H.C.W. Lau, S.Y. Chung, and T. Tsoi. 2021. American sign language recognition and training method with recurrent neural network. *Expert Systems with Applications*, 167(114403).

C.K. Mummadi, F.P.P. Leo, K.D. Verma, S. Kasireddy, P.M. Scholl, J. Kempfle, and K.V. Laerhoven. 2018. Real-time and embedded detection of hand gestures with an imu-based glove. *Informatics*, 5(2):1–18.

C. Neidle, A. Thangali, and S. Sclaroff. 2012. Challenges in development of the american sign language lexicon video dataset (asllvd) corpus. In *In Proceedings of the 5th Workshop on the Representation and Processing of Sign Languages:*

Interactions between Corpus and Lexicon, pages 1–8, Online.

Z. Niu and B. Mak. 2020. Stochastic fine-grained labeling of multi-state sign glosses for continuous sign language recognition. In *In Proceedings of the European Conference on Computer Vision (ECCVâĂŹ20)*, pages 172–186, Glasgow, United Kingdom.

M. Oszust and M.J. Wysocki. 2013. Polish sign language words recognition with kinect. In *In Proceedings of the 6th International Conference on Human System Interactions (HSI'13)*, pages 219–226, Gdansk, Poland.

N. Pugeault and R. Bowden. 2011. Spelling it out: real-time asl finger spelling recognition. In *In Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1114–1119, Barcelona, Spain.

F. Ronchetti, F. Quiroga, C. Estrebou, L. Lanzarini, and A. Rosete. 2016. Lsa64: an argentinian sign language dataset. In *In Proceedings of the 22nd Congreso Argentino de Ciencias de la Computación (CACIC'16)*, pages 794–803, San Luis, Argentina.

S.M. Shohieb, H.K. Elminir, and A.M. Riad. 2015. Signsworld atlas; a benchmark arabic sign language database. *Journal of King Saud University-Computer and Information Sciences*, 27(1):68–76.

Y.H. Sun, N. Kuwahara, and K. Morimoto. 2013. Development of recognition system of japanese sign language using 3d image sensor. In *In Proceedings of the International Conference on Human-Computer Interaction (HCII'13)*, pages 286–290, Toronto, Canada.

A. Tharwat, T. Gaber, A.E. Hassenian, M. K. Shahin, and B. Refaat. 2015. Sift-based arabic sign language recognition system. In *In Proceedings of the Afro-European Conference for Industrial Advancement (AECIA'15)*, pages 359–370, Villejuif, France.

S. Tornay, M. Razavi, and M.M. Doss. 2020. Towards multilingual sign language recognition. In *In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICAASP'20)*, pages 6309–6313, Online.

D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *In Proceedings of the IEEE International Conference on Computer Vision (ICCV'15)*, pages 4489–4497, Santiago, Chile.

X.Y. Wang, M. Xia, H.W. Cai, Gao Y, and C. Cattani. 2012. Hidden-markov-models-based dynamic hand gesture recognition. *Mathematical Problems in Engineering*, pages 1–11.

S.H. Yang, S.J. Jung, H.W. Kang, and C. Kim. 2020. The koren sign language dataset for action recognition. In *In Proceedings of the International Conference on Multimedia Modeling (MMM'17)*, pages 532–542, Reykjavik, Iceland.

F. Yin, X.J. Chai, Y. Zhou, and X.L. Chen. 2015. Semantics constrained dictionary learning for signer-independent sign language recognition. In *In Proceedings of the IEEE International Conference on Image Processing (ICIP'15)*, pages 3310–3314, Quebec City, Canada.

T. Zhang, W.M. Zheng, Z. Cui, Y. Zong, C.L. Li, X.Y. Zhou, and J. Yang. 2020. Deep manifold-to-manifold transforming network for skeleton-based action recognition. *IEEE transactions on multimedia*, 22(11):2926–2937.