Multimedia Tools and Applications (2024) 83:50245–50267
https://doi.org/10.1007/s11042-023-17268-8

Diverse hand gesture recognition dataset

Zahra Mohammadi[1] · Alireza Akhavanpour[2] · Razieh Rastgoo[3] · Mohammad Sabokrou[4]

Received: 25 July 2022 / Revised: 18 May 2023 / Accepted: 22 September 2023 /
Published online: 4 November 2023

Abstract

Hand Gesture Recognition(HGR) is a challenging computer vision task. Recently, by taking advantages of deep learning-based models, HGR methods have achieved outstanding results and outperformed state-of-the-art alternatives by a high margin. However, the performance of deep learning-based models is highly dependent on the data. A large amount of data is required to train deep learning-based models. While there are some widely-used datasets in HGR, these datasets lack diverse gestures in real-world situations. To this end, we propose a hand gesture dataset (Dataset will be publicly available after paper publication.), including diverse gestures with more sample numbers per gesture class. Furthermore, we provide hand annotations, including a hand bounding box, 3D hand keypoints, and gesture label per sample. The proposed dataset aims to provide a benchmark for research works to tackle real-world situations. The dataset samples are recorded in a real-world background with high complexity and diversity. To be more realistic, the proposed dataset does not include any pre-processing step. All of the samples in this dataset are pure and real. This configuration makes room to underpin future research works in a real-world situation and develop gesture recognition models in an unrestricted environment. Overall, our dataset outperforms in terms of diversity, number of subjects, number of samples per gesture class, and use of real data. Finally, different analysis on the existing state-of-the-art models in HGR, HPE, hand recovery, and hand reconstruction were performed and reported. Our implementation is available at https://github.com/smohammadi96/Diverse_hand_gesture_dataset/blob/main/README.md.

✉ Razieh Rastgoo
rrastgoo@semnan.ac.ir

Zahra Mohammadi
s.mohammadi@shenasa.ai

Alireza Akhavanpour
akhavan@shenasa.ai

Mohammad Sabokrou
sabokro@ipm.ir

1 Computer Science and Engineering Department, Shahid Beheshti University, Tehran 1983969411, Iran

2 Artificial Intelligence group, Shenasa, Tehran 1456755131, Iran

3 Electrical and Computer Engineering Department, Semnan University, Semnan 3513119111, Iran

4 Institute for Research in Fundamental Sciences (IPM), Tehran 193955746, Iran

Keywords Hand Gesture Recognition (HGR) · Deep learning · Dataset · Hand Pose Estimation (HPE) · Computer vision

1 Introduction

Communication among people has various forms, including verbal and non-verbal. While verbal communication is mainly relying on spoken language, non-verbal communication uses body language or gestures. The hand, as one of the main body parts in communication, can effectively convey information from one person to another. To facilitate communication between different groups of people, hand gestures play a key role in both verbal and non-verbal languages. To this end, proposing a hand gesture-based system is in line with this requirement [61, 96, 97].

Hand Gesture Recognition(HGR), an interesting sub-field within computer vision, has recently received high attention. Hand gesture is critical for human behavior understanding focusing on recognizing more fine-grained upper body movements within the special context. One of the earliest works for HGR [35, 89] was based on traditional hand-crafted features. In the past decades, this task as well as all other computer vision tasks have been revolutionized by the advent of deep learning and achieved State-Of-The-Art (SOTA) performances [23, 45, 61, 74, 86]. The HGR can be used in different computer vision tasks such as sign language recognition [62, 64–66], security [83], e-health [27, 83], and entertainment [47].

Generally, HGR models can be categorized into two main categories: static and dynamic. The static recognition methods only consider spatial features obtained from the input images while the dynamic methods benefit from both spatial and temporal features extracted from videos. One of the main challenges in both categories is the dataset. Different characteristics are considered for the dataset evaluation, such as the number of gestures, the number of samples in each gesture, the number of subjects, and the diversity of the samples. Various datasets have been developed to meet these characteristics. Considering the application domain, each characteristic can have a decisive role. Generally, deep learning-based models need datasets including a large number of samples in each gesture. However, increasing the diversity of the samples can help the model to work in real-world situations. In line with requirements, we propose a dataset for the static HGR task which contains more samples in each class captured in more diverse conditions. No pre-processing is performed on the images of this dataset aiming to be usable as a baseline in real-world conditions.

The remainder of this paper is organized as follows. Section 2 presents a brief review of related works in gesture recognition. The existing hand gesture datasets are briefly presented in Section 3. Details of the proposed dataset are introduced in Section 4. Then, Section 5 presents some experimental results on the proposed dataset. Limitations of the collected dataset are also discussed in Section 6. Finally, we conclude the work in Section 7.

2 Related work

Here, we briefly review recent work and datasets in HGR.

2.1 Deep learning-based HGR

Hand gesture, containing complementary information, plays an important role in better communication in daily communication. Due to various applications of HGR, researchers have focused on this task and several methods have been proposed [27, 47, 63, 83, 93–95]. HGR is one of the main components of Sign Language Recognition (SLR), whereas facial expressions and body actions play the role of giving emphasis to the words and phrases conveyed by hand gestures. Generally, deep learning-based HGR models can be categorized into static HGR and dynamic HGR models:

- **Static HGR** Static HGR is mainly focusing on the various shapes and orientations of hands without considering the motion information. Rautaray and Agrawal developed a hand gesture recognition system for interacting with different applications to present an applicable solution towards a handy interface between human and computer. This system employs image processing techniques for detection, segmentation, tracking, and recognition of hand gestures for converting it to a meaningful command. Furthermore, this system can be used for controlling different applications like game control [67]. Ameen et al. proposed a model using a Convolutional Neural Network (CNN) model for static HGR from letters of the American Sign Language (ASL) alphabet. Two modalities, RGB and Depth, are used in parallel to obtain the spatial features from two CNNs. Results on the ASL fingerspelling dataset show the recognition accuracy of 80.34% [4]. Rastgoo et al. proposed a Restricted Boltzmann Machine (RBM) and CNN-based model to recognize the ASL letters from two input modalities,RGB and Depth. A CNN model is used for hand detection. Results on four public datasets, Massey University Gesture Dataset, ASL, and Fingerspelling Dataset from the University of Surrey's Center for Vision, Speech and Signal Processing, NYU, and ASL Fingerspelling A, confirm the superiority of the model performance with a relative accuracy improvement of 27.31% 28.56% 2.9% and 11.13% respectively [62]. Mohanty et al. suggested a CNN-based model for static HGR from the complex background and varying illumination conditions in the input images. Results on three publicly available datasets, the NUS hand posture dataset with a cluttered background, Triesh hand posture dataset with the uniform dark background, and Marcel hand posture, and obtained comparable results with SOTA models [51]. Adithya and Rajesh proposed a CNN model for static HGR from RGB images. The proposed method has been evaluated on two public datasets, NUS hand posture and American fingerspelling A, using five-fold cross-validation. Results show that the model has comparable results with SOTA models in the field [2]. Moghbeli Damaneh et al. designed a model, including a CNN as well as a classical non-intelligent feature extraction method, for static hand gesture recognition. After preprocessing and removing the image background, it passes through three different streams of feature extraction, containing the CNN, the Gabor filter, and ORB feature descriptor. Finally, all features obtained from three streams are fused and fed to the final classifier. Results on three public datasets show the promising results of the model [49]. John and Deshpande suggested a deep learning-based model, entitled Multi-Dilated Convolution-based DenseNet (MDCDN), including a combination of multi-dilated convolution and DenseNet. Results show the efficiency of deep features in accurate HGR [30]. Mohammadi et al. designed four Spiking Neural Network (SNN) models for two static American Sign Language (ASL) hand gesture classification tasks: the ASL alphabet and ASL digits. The compared the SNN to the equivalent Deep Neural Network (DNN) models and reported the results in terms of accuracy, latency, power consumption, and energy. In overall, the best DNN model had a higher performance than the SNN models [50]. Yu et al. suggested a model using two input modalities, the RGB and optical flow keyframes, for dynamic gesture recognition. The spatial and temporal features are extracted using a 2D CNN and fused for final classification. Results on two public datasets, Cambridge Hand Gesture dataset and Northwestern University Hand

Gesture dataset, show the efficiency of the model in term of recognition accuracy [90]. Sharma and Singh developed a CNN-based model for HGR in sign language. Different CNN architectures, such as VGG-11 and VGG-16, have been evaluated on two public datasets. Furthermore, a dataset containing 2150 RGB images of Indian Sign Language (ISL) gestures has been collected. Results show the promising results of this model as well as being invariant to rotation and scaling transformation [73].

- **Dynamic HGR** Dynamic HGR is working on the sequence of hand postures with associated motion information. Roccetti et al. designed and implemented a multimedia system to mimic the process of cooking. In this system, a virtual experience is focused on simulating the movements a real cook to prepare its recipe. The main contribution of this model is recognizing and tracking the actions and gestures during the cooking [68]. Roccetti et al. discussed the differences between the design of a gesture-based interface for a console as well as a similar one for a public space setting for gaming. In addition, they have employed a set of algorithms that have specifically designed for gesture-based interfaces for public spaces. Results showed that relying on a video camera and a robust gesture recognition software system, their model obtained a promising performance [69]. Elboushaki et al. proposed a multi-dimensional CNN model for dynamic HGR from RGB-D videos. A combination of 3DCNN and LSTM is used for spatio-temporal feature extraction. Results of the model on three datasets, SKIG, NATOPS, and SBU, show a relative recognition accuracy improvement of 0.19% 8.52% and 4.11% compared to SOTA models [20]. Chen et al. suggested a Dynamic Graph-based Spatial-Temporal Attention (DG-STA) model for dynamic HGR. Using a self-attention mechanism for learning the node and edge features, a graph is constructed from a hand skeleton. Results on two datasets, DHG-14/28 and SHREC'17, show the recognition accuracy improvement with a 0.9% and 3% relative improvement, respectively [13]. Canuto dos Santos et al. proposed a CNN-based model by using a soft-attention layer for dynamic HGR. They employed a summarizing technique to obtain an RGB image from an RGB video. This image is fed to the CNN-based model to obtain the final gesture. Results on Montalbano and GRIT datasets show a relative SOTA accuracy improvement of 0.78% and 6.68% respectively [18]. Subhashini and Revathi present a Gabor Line Derivative Deep Convolution Neural Network-based Levy flight Whale optimization for static and dynamic hand gesture recognition. After decreasing the computation complexity of the image channels, a rich set of line features are extracted using the Gabor Line Derivative-based feature extraction method. Finally, a Deep Convolution Neural Network based Levy flight Whale optimization is presented in terms of classifying dissimilar static and dynamic hand gestures. Results of this model confirm the superiority of the model in compared to state-of-the-art models in HGR [77].

3 HGR datasets

Recently, several datasets have been proposed to HGR with various characteristics [6, 8, 21, 36, 58, 60, 63, 82]. In this section, we briefly review these datasets from four characteristics points of view. Table 1 shows the most used datasets for gesture recognition.

- **The gesture numbers** There are different gesture numbers, from 10 to 3300, in the gesture datasets. As the fourth column of Table 1 shows, the Boston ASLVID dataset has the highest gesture numbers among the current gesture datasets. While increasing the gesture numbers has many advantages for a model, it is not enough for a deep learning-based model.

Table 1 The most used emblems datasets

Y	Dataset	C	CN	SubN	SampN	SPC	LL	A	Ava.
2011	Boston ASL LVD [81]	USA	3300	6	9800	3	W	H	P
2012	DGS Kinect 40 [14]	Germany	40	15	3000	8	W	–	P
2012	RWTH-PHOENIX-Weather [32]	Germany	1200	7	45760	5	S	F, H	P
2012	GSL 20 [57]	Greek	20	6	840	5	W	–	P
2012	PSL Kinect 30 [58]	Poland	30	1	300	10	W	–	P
2013	PSL ToF 84 [58]	Poland	84	1	1680	10	W	–	P
2014	DEVISIGN-G [38]	China	36	8	432	5	W	–	P
2014	DEVISIGN-D [38]	China	500	8	6000	5	W	–	P
2014	DEVISIGN-L [38]	China	239	8	24000	5	W	–	P
2015	SIGNUM [32]	Germany	455	25	33210	3	S	–	P
2016	MSRGesture [85]	USA	12	10	336	3	W	–	P
2016	LSA64 [71]	Argentina	64	10	3200	5	W	H, h	P
2016	TVC-hand gesture [31]	Korea	10	1	650	5	–	–	P
2020	RKS-PERSIANSIGN [63]	Iran	100	10	10000	100	W	H	P
2021	**DiverseHandGesture**	**USA**	**8**	**49**	**7990**	**800**	**W**	**H**	**WP**

Y: Year, C: Country, CN: Class Number, SubN: Subject Number, SampN: Sample Number, SPC: Sample Per Class, LL: Language Level (word or sentence), A: Annotation, F: face, H: hand, h: head, W: word, S: sentence, Ava.: Availability

- **The sample numbers in each gesture** As the seventh column of Table 1 shows, RKS-PERSIANSIGN with 100 samples has the highest sample numbers. The Boston ASLVID, SIGNUM, and MSRGesture have 3 samples for each gesture. Increasing the sample numbers is necessary for deep learning-based models.
- **The sample numbers** Increasing the sample numbers can help the model to face more data patterns and learn to be more general. As the sixth column of Table 1 shows, the RWTH-PHOENIX-Weather dataset with 45760 samples and PSL Kinect 30 with 300 samples have the highest and lowest sample numbers, respectively. However, having more sample numbers is not enough for model learning.
- **The subject numbers** Increasing the subject numbers can assist the model to be subject-independent and work in real-world situations, including diverse samples. This can lead to an increase in the generalization capability of the model. As the fifth column of Table 1 shows, there are different subject numbers, from 1 to 25, in the gesture datasets.

Considering these characteristics, we propose a dataset, including more sample numbers in each gesture and also more subject numbers. As the last row of Table 1 shows, the proposed dataset contains 800 samples in each gesture and also 49 subject numbers. Relying on these characteristics of the proposed dataset, deep learning-based models can learn each gesture with a large amount of samples in diverse situations. However, we include only 8 gestures, corresponding to the most-used gestures in daily communication. Our contribution will be extended to include more gestures in our dataset.

4 Proposed dataset

Here, we present the details of the proposed dataset.

4.1 Dataset overview

The dataset contains a total of 7990 RGB images belonging to 8 most-used gesture categories in daily/social communication, in different and complex backgrounds, performed by 49 actors. We have 37 individuals, including 22 men and 15 women, in the age interval of 18-45 for the training set. There are 12 individuals, including 6 men and 6 women, in the age interval of 23-35 for the testing set. Furthermore, the training and testing sets contain a total of 6400 and 1590 samples, respectively. This dataset was collected by a webcam in PNG format with medium quality in different lighting conditions and background complexity. The distance configuration between an actor and the camera is diverse. The input shape of the dataset samples is $224 \times 224 \times 3$. The gesture classes are chosen from the most usable and functional gestures used in daily communication. To develop a set of useful gestures for social media or face-to-face meetings, we began by identifying twenty potential gestures. We then solicited feedback from a diverse group of individuals to determine which gestures were most useful and practical. After careful consideration and voting, we narrowed down the list to eight gestures that are both highly effective and easy to perform. These gestures were selected based on their potential to enhance communication and facilitate understanding in a variety of settings. Table 2 shows a description of the gestures used in the proposed dataset. We used Labellmg software [37] to manually annotate the RGB images of the proposed dataset. The annotation includes the bounding box of the detected hand, the class label corresponding to the gesture of the input image, and the hand pose parameters. The proposed dataset aims to

Table 2 Details of the gestures used in the proposed dataset

Gesture	Description
"exactly/OK"	This gesture is used when we agree with an idea or suggestion.
"two"	This gesture is a number and can be used to show success in sign language.
"five"	This gesture is a number and can be used as a stopping action in sign language.
"left"	This gesture shows the orientation in directions.
"three"	This gesture is a gesture that people around the world have made for centuries, mostly in positive contexts. It is used for several purposes in sign languages, and in yoga as a symbol to demonstrate inner perfection.
"like"	This gesture is used when we agree and like something.
"dislike"	This gesture is used when we disagree and don't like something.
"zero"	This gesture is a number and shows a so-so reaction to something.

provide a benchmark for HGR in real-world conditions. Our dataset will be publicly available in the future.

4.2 Demo and repository

We develop an API [7][1] along with a GitHub repository[2] for real-time HGR. The most used gestures in daily communication are included in this system. CNN, as a powerful deep learning-based model, is the main core of this system. This web service can be used in Sign Language Recognition and authentication systems to solve communication barriers. Figure 1 shows an overview of the developed API.

4.3 Dataset statistics

Here, we present a statistical overview of the proposed dataset, including the gesture numbers, sample numbers per gesture class, background complexity, subject diversity, age diversity, hand occlusion, and hand pose parameters.

- **Gesture numbers** After reviewing the most-used gestures in daily communication, we included 8 gestures in the dataset.
- **Sample numbers per gesture class** There are a total of 800 samples per gesture label, which is compatible with the prerequisites of deep learning-based models.
- **Background complexity** To make a real-world recognition, different backgrounds, including different complexity levels, are included in the dataset.
- **Subject diversity** To make a person-independent benchmark, different subjects with different configurations are presented in the dataset.
- **Age diversity** To include different hand scales and shape configurations in the dataset, subjects in the dataset fall into different age groups (See Table 3).
- **Hand occlusion** To have a robust benchmark, gestures with different hand occlusions are included in the dataset.
- **Hand pose parameters** We include 21 3D hand keypoints for each sample in the dataset.

[1] This demo is available at http://shenasa.ai/service/59/hand-gesture-recognition

[2] Will be available at https://github.com/smohammadi96/Diverse_hand_gesture_dataset/blob/main/README.md

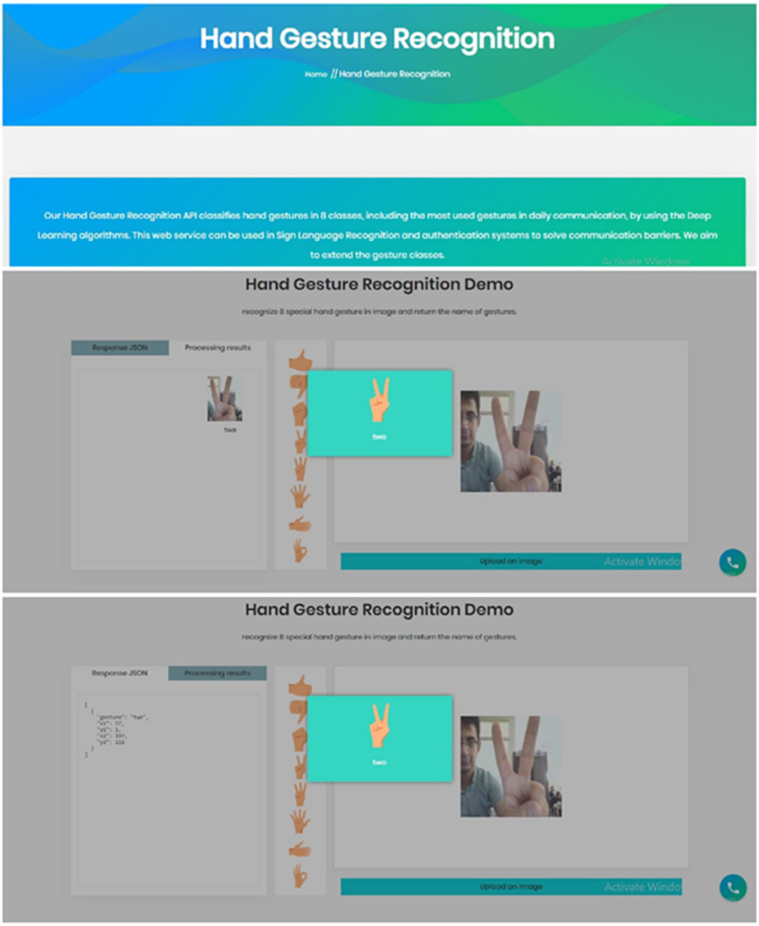

Fig. 1 An overview of the proposed API

Details of the proposed dataset, along with some samples, are shown in Table 3 and Fig. 2.

5 Evaluated algorithms and baselines

Hand gesture recognition aims to map the hand's appearance and/or motion related features to a gesture vocabulary set. During this mapping, some features, such as 2D/3D hand pose and shape features, can be used. However, estimating the 3D hand pose features is a challenging

Table 3 Details of the proposed dataset

	Number of men	Number of women	Age Range Women	Age Range men	Total images
Train set	22	15	18-30	22-45	6400
Test set	6	6	23-28	23-35	1590

Fig. 2 Some samples of the proposed dataset

task, especially in RGB image/videos. In this way, hand recovery can be used in coping with the occluded or damaged input. Due to the relation between the hand gesture recognition, hand pose estimation, and hand recovery, we briefly discuss recent works in these areas in this section.

5.1 HGR

Here, we report some experimental results regarding HGR performed on the proposed dataset. Some of the CNN models used in Tables 4 and 5 are fine-tuned to be compatible with the proposed dataset. Due to the undeniable power of CNN models for feature extraction from static images, we use some CNN-based for analysis. One of these CNN-based models is DeepGesture [3] as a deep learning-based model, including an Adapted Deep Convolutional

Table 4 Experimental results on the proposed dataset

Model name	Include Top	Accuracy
DeepGesture [3]	–	0.95
DenseNet121 [29]	True	0.97
inceptionv3 [79]	True	0.95
NASNetLarge [92]	True	0.95
ResNet50 [26]	True	0.98
ResNet50 [26]	False	0.97
VGG16 [75]	True	0.97
inception-resnet-v2 [78]	True	0.87
MobileNet [28]	True	0.94
Xception [88]	True	0.94
MobileNetv2 [72]	True	0.94

Table 5 Results of some CNN-based models on the proposed dataset

class name	EfficientDet0	YOLOv3-tiny	EfficientDet1	EfficientDet2	EfficientDet3	EfficientDet4	YOLOv4
Exactly	0.94	0.94	0.93	0.93	0.88	0.89	0.91
Five	0.90	0.86	0.93	0.88	0.89	0.92	0.98
Two	0.87	0.86	0.79	0.92	0.87	0.88	0.94
Three	0.71	0.76	0.86	0.85	0.77	0.76	0.90
Zero	0.94	0.95	0.95	0.95	0.93	0.96	0.96
Left	0.99	0.97	0.99	0.98	0.99	0.98	0.99
Like	0.84	0.92	0.95	0.90	0.93	0.90	0.98
Dislike	0.98	0.93	0.99	0.98	0.92	0.83	1.00
Total accuracy	0.89	0.89	0.92	0.92	0.89	0.88	**0.95**

Neural Network (ADCNN), proposed to HGR. Two convolutional and three Fully Connected (FC) layers are included in this model. Table 4 shows the results of this model on the proposed dataset. Furthermore, results on some CNN-based models, including DenseNet121 [29], inceptionv3 [79], NASNetLarge [92], ResNet50 [26], VGG16 [75], Inception-ResNet-v2 [78], MobileNet [28], Xception [88], and MobileNetv2 [72], are reported in the Table 5. Finally, we report the results of three well-known real-time object detection and classification models: YOLOv3-tiny [1], YOLOv4 [9], and different architectures of EfficientDet [80] (see Table 5). As this table shows, YOLOv4 has a better recognition accuracy compared to the other models used in the evaluation. While the recognition accuracy is 100% for the “dislike” gesture using the YOLOv4 model, the “three” gesture is challenging in all models used in the evaluation. The “left” gesture has an approximately stable behavior for all models. Due to the high sample numbers per class and also using seen data for testing, the results of these tables are satisfactory. Furthermore, Table 6 shows the statistical evaluation of the models. Based on this table, YOLOv4 has a minimum Standard Deviation (std) compared to the other models.

5.2 Hand and body pose estimation

The proposed dataset is collected for HGR but it can be used in different tasks in Computer Vision. We performed different analysis for hand and body pose estimation on the proposed dataset. One of them is the accurate Hand Pose Estimation (HPE) model proposed by Zimmermann and Brox [91]. This model contains a CNN-based model for hand segmentation and localization to estimate 21 3D hand keypoints from RGB images. We did not train this model and only used it as a pre-trained model for 3D hand keypoints estimation. Figures 3 and 4 show the results of the estimated hand keypoints on different samples. As these figures show, while the Zimmermann and Brox [91] model can successfully estimate 3D hand keypoints in some samples (Fig. 3), the estimation accuracy is decreased under the effect of lighting conditions and hand occlusion in some samples (Fig. 4). Another model is the OpenPose model, as the state-of-the-art model in pose estimation [10]. The first row of Fig. 5 shows some outputs of this model. Furthermore, we did some experiments on the proposed datasets for hand and body pose estimation models proposed in [11, 42, 87] (See the first, second, and fourth rows of Fig. 6). The first row of Fig. 6 shows the results of the multi-person pose estimation model [42]. As this figure shows, the estimation is not accurate that can come from the complexity of the background and also light conditions in dataset samples.

Table 6 Statistical evaluation of the considered models for evaluation on the proposed dataset

Model name	Mean	Std
EfficientDet0	0.31	0.31
EfficientDet1	0.64	0.45
EfficientDet2	1.31	1.06
EfficientDet3	1.7	0.56
EfficientDet4	4.13	1.20
YOLOv3-tiny	0.10	0.05
YOLOv4	0.27	**0.04**

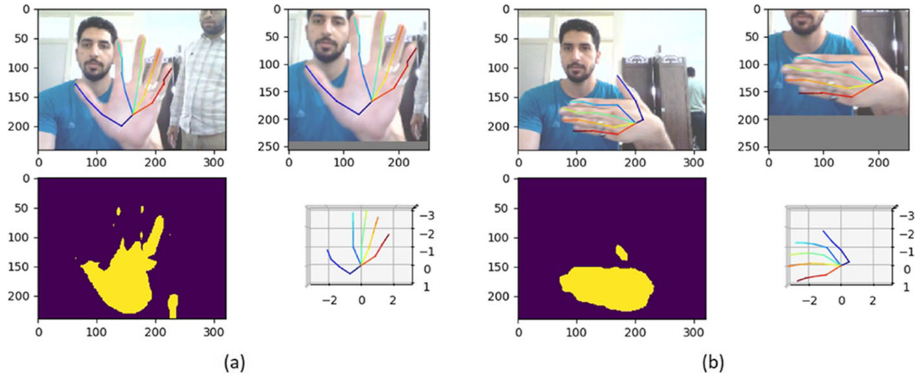

Fig. 3 Some samples of accurate 3d hand keypoints estimated using Zimmermann and Brox [91] model on the dataset samples

5.3 Hand recovery

3D hand shape recovery is a challenging task in computer vision. In 3D capturing the full part of the body, the hands are small and sometimes partially occluded and damaged. So, in some cases, it is necessary to recover and reconstruct their shape [70]. This process is defined as hand recovery. Here, we present the results of the proposed dataset on the HandTailor model, as a high-precision hand recovery model. This model combines a learning-based hand module and an optimization-based tailor module to obtain an accurate hand mesh recovery from an RGB image [43]. As the third row of Fig. 6 shows, this model does not have accurate results on some samples from the proposed dataset.

5.4 Joint and mesh reconstruction of hands

Estimating hand-object manipulations is necessary for interpreting and imitating human actions. Due to this importance, we performed an analysis on a model, entitled ObMan, for

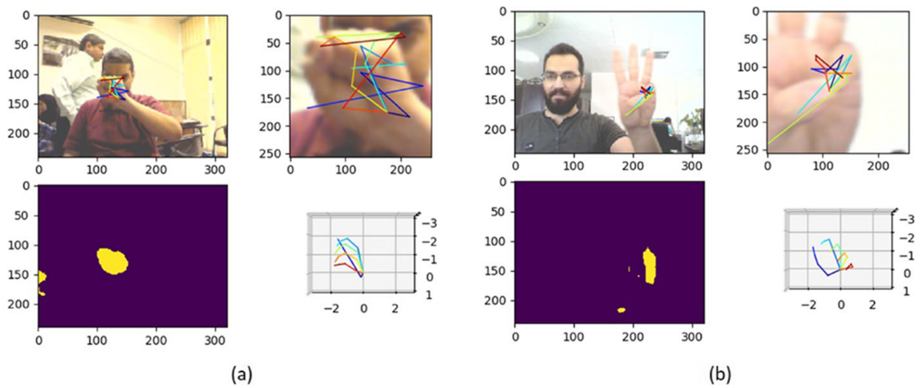

Fig. 4 Some samples of inaccurate 3d hand keypoints using Zimmermann and Brox [91] model on the dataset samples

the joint reconstruction of hands and objects from an RGB image. The ObMan regularizes the joint reconstruction of hands and objects with manipulation constraints [25]. As the fifth row of Fig. 6 shows, this model does not have accurate results on some samples from the proposed dataset. Furthermore, the results of two models [41, 59] for hand mesh estimation and reconstruction have been shown in the third and fourth rows of Fig. 5.

5.5 State-of-the-art evaluation

Here, we present aggregated information about the state-of-the-art results on HGR and related areas (See Tables 7 and 8). As these tables show, trends of the proposed models on different datasets in static and dynamic HGR and related areas show that deep learning approaches successfully improved the model performance with a high margin. However, more endeavor is necessary for some challenging datasets such as isoGD, LSP, and EVAL. In most of the existing datasets, such as NYU, ICVL, MSRA, ASL Fingerspelling A, RKS-PERSIANSIGN, the achieved performance by deep-based models are higher than the other challenging datasets. The proposed experimental results of different deep-based models in static and dynamic HGR and related areas confirm the effective role of using multi-modal and multi-channel information [19, 20, 62, 76]. Furthermore, the proposed hybrid models successfully improved the model performance benefiting from the combination of some hand-crafted features with

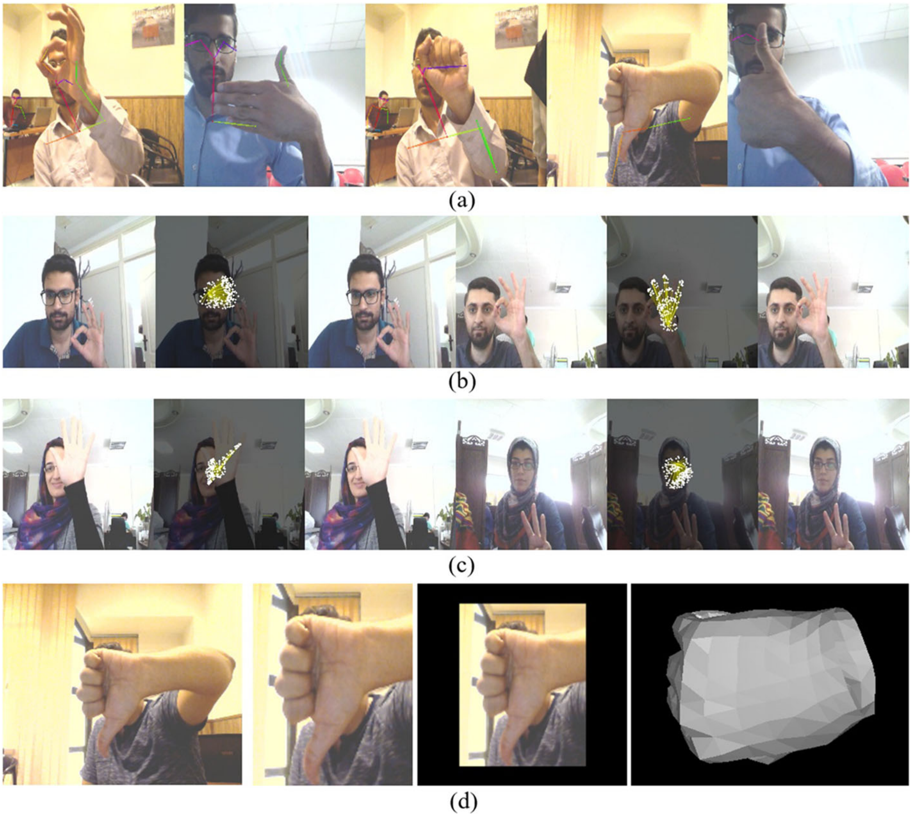

Fig. 5 Some experiments of the proposed datasets: (a) [10] (b) and (c) [41], (d) [59]

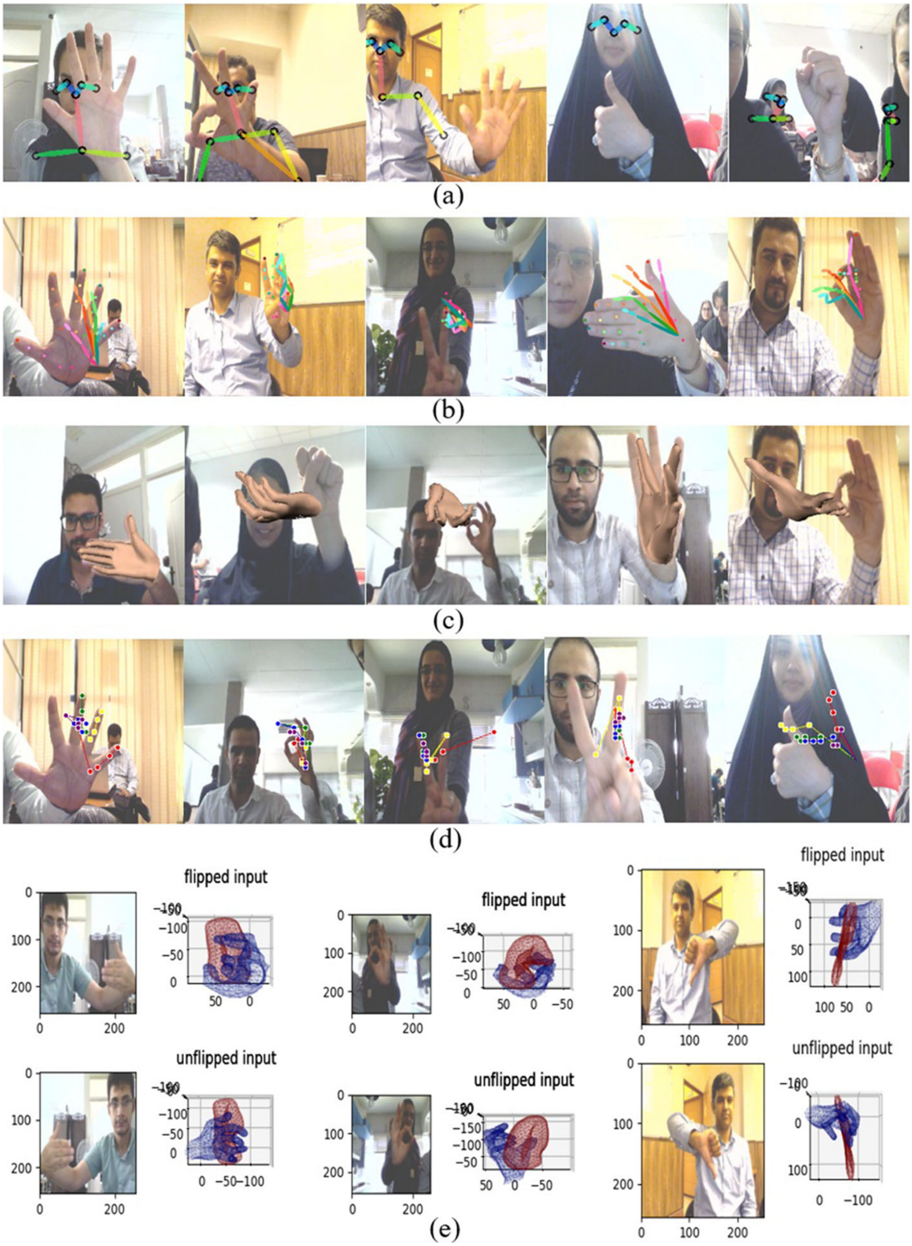

Fig. 6 Some experiments of the proposed datasets: (a) [42] (b) [87] (c) [43] (d) [11] (e) [25]

Table 7 State-of-the-art models on the datasets corresponding to the HGR and related areas

Dataset	Year	Ref.	Goal	Model	Modality	Results
NYU	2020	[63]	HSR	SSD, 2DCNN, 3DCNN, LSTM	Depth	4.64 mm
ICVL	2018	[52]	HP	CNN	Depth	6.28 mm
MSRA	2016	[56]	HP	CNN	Depth	5.58 mm (Ave. err.)
FLIC	2016	[54]	HP	CNN	RGB	99.0 (Elbow)
LSP	**2016**	[87]	**HP**	**CNN**	**RGB**	**84.32**
isoGD	2020	[64]	HSR	SSD, CNN, LSTM	RGB	86.32
MPII	**2016**	[54]	**HP**	**CNN**	**RGB**	**90.90 (total)**
ITOP	2018	[46]	HP	CNN	Depth	97.5 (AUC)
RGBD-HuDaAct	2016	[19]	HG	CNN	Depth, RGB	96.74
STB	**2018**	[76]	**HP**	**VAE**	**RGB, Depth**	**0.983(AUC)**
EVAL	2016	[24]	HP	CNN	Depth	74.10
Dexter	**2017**	[91]	**HP**	**CNN**	**RGB**	**49.0 (AUC)**
RWTH-PHOENIX-Weather 2012	2015	[33]	HSR	CNN	RGB	55.70 (Precision)
RWTH-PHOENIX-Weather 2014	2019	[15]	CDSLR	3DCNN, Bi-LSTM	RGB	22.86
BigHand2.2M	2018	[5]	HP	GAN	Depth	13.7 mm
Human3.6M	2018	[84]	HP	CNN	Depth	62.8 mm
NGT	2017	[48]	CDSLR	heuristic, LSTM	RGB	80.70 (accuracy)
UBC3V	2018	[46]	HP	CNN	Depth	88.2 (AUC)
Massey 2012	**2018**	[62]	**HR**	**RBM**	**RGB, Depth**	**99.31**
SL Surrey	2018	[62]	HR	RBM	RGB, Depth	97.56
ASL Fingerspelling A	**2018**	[62]	**HR**	**RBM**	**RGB, Depth**	**98.13**
OUHANDS	2018	[16]	HG	CNN	Depth	86.46
Egohands	**2017**	[17]	**HT**	**CNN**	**RGB**	**0.9686 (mAP)**

Table 7 continued

Dataset	Year	Ref.	Goal	Model	Modality	Results
Dexter	2018	[53]	HT	CNN	RGB	0.64 (AUC)
EgoDexter	2018	[53]	HT	CNN	RGB	0.54 (AUC)
RHD	**2018**	[76]	**HP**	**VAE**	**RGB, Depth**	**0.849(AUC)**
B2RGB-SH	2019	[39]	HP	CNN	RGB	7.18 (err)
DHG-14/28 Dataset	2019	[13]	HG	CNN	RGB	91.9
SHREC'17 Track Dataset	2019	[13]	HG	CNN	RGB	94.4

The results of the static hand gesture datasets are shown in bold

Table 8 State-of-the-art models on the datasets corresponding to the HGR and related areas

Dataset	Year	Ref.	Goal	Model	Modality	Results
RWTH-BOSTON-50	2019	[40]	HS	CNN	RGB	89.33
ASLLVD	2019	[40]	HS	CNN	RGB	31.50
EgoGesture	**2019**	[34]	**HG**	**CNN**	**RGB**	**94.03**
NVIDIA benchmarks	2019	[34]	HG	CNN	RGB	83.83
SKIG	2020	[20]	HG	CNN	RGB, Depth	99.72
NATOPS	2020	[20]	HG	CNN	RGB, Depth	95.87
SBU	2020	[20]	HG	CNN	RGB, Depth	97.51
NUS	**2020**	[2]	**HG**	**CNN**	**RGB**	**94.7**
First-Person	2020	[63]	HSR	SSD, 2DCNN, 3DCNN, LSTM	RGB	91.12
RKS-PERSIANSIGN	2020	[63]	HSR	SSD, 2DCNN, 3DCNN, LSTM	RGB	99.80

Results of the static hand gesture datasets are shown in bold

deep-based features [12, 22, 44, 64]. These models benefit from having a trade-off between both the powerful capabilities of deep learning (in particular in those cases of having large amounts of data) and the specific problem-tailored design of handcrafted features. Due to the undeniable power of CNN models for feature extraction from visual inputs, in most of the proposed deep-based models, CNN or a combination of CNN with other deep-based models is employed. Generative models, such as RBM and VAE, showed a comparable or better performance than other deep alternatives in coping with few data for HGR and related areas [62, 76]. Since the dynamic modality is more challenging than the static one, most of the proposed models employed LSTM or 3DCNN for analyzing temporal dynamics. Considering the scope of the proposed dataset in this work, state-of-the-art results on static datasets have been discussed in these tables. Most of the proposed models benefit from the powerful capabilities of CNN for spatial feature extraction from static images. However, VAE and RBM are also used. Overall, the model performance of the models on some complex static and dynamic datasets can be improved. A compact details of the state-of-the-art models used for evaluation can be found in Table 9. Finally, to show the complexity of the proposed dataset, we compare the results of some recent models for the available datasets as well as the proposed dataset (See Table 10). As this table shows, results on the proposed dataset are lower than the available datasets in these works. This comes from the higher complexity and diversity of the proposed dataset.

6 Limitations

The hand gesture dataset we have proposed is suitable for static gesture recognition. However, if we collect sequential frames, we can use them for action recognition as well. It is important to note that the number of classes in the dataset can be expanded in the future to increase its versatility. Additionally, the quality of the images captured by a laptop's camera is limited, and it is recommended to use different cameras and lenses to capture images with better variation and quality. By addressing these limitations, we can improve the accuracy and usefulness of the dataset for a wider range of applications.

Table 9 Experimental results on the proposed dataset

Ref.	SOTA title	Task	year	link
[42]	SimplePose	multi-person pose estimation	AAAI 2020	https://github.com/hellojialee/Improved-Body-Parts
[87]	Convolutional Pose Machines	Hand and body Pose estimation	CVPR 2016	https://github.com/timctho/convolutional-pose-machines-tensorflow
[59]	HandOccNet	3D mesh estimation	CVPR 2022	https://github.com/namepllet/HandOccNet
[43]	HandTailor	3D Hand Recovery	BMVC 2021	https://github.com/LyuJ1998/HandTailor
[41]	MeshTransformer	Hand pose and mesh reconstruction	CVPR 2021	https://github.com/microsoft/MeshTransformer
[11]	NSRMhand	2D hand pose estimation	WACV 2020	https://github.com/HowieMa/NSRMhand
[25]	Obman	Joint Reconstruction of Hands and Manipulated Objects	CVPR 2019	https://github.com/hassony2/obman-train
[10]	Openpose	Hand and body pose estimation	PAMI 2018	https://cmu-perceptual-computing-lab.github.io/openpose/web/html/doc/md-doc-installation-0-index.html

7 Conclusion

In this paper, we contributed to a dataset including a total of 7990 RGB images in 8 gesture classes. The proposed dataset aimed to include a diverse set of hand gesture images in the most used gestures corresponding to daily communication. Analysis of the existing datasets in HGR

Table 10 The comparison results of some models for the available datasets with the proposed dataset

Model	Own dataset [55]	Massey [62]	Surrey [62]	NYU [62]	Fingerspelling A [62]	Proposed dataset
Decision Tree [55]	91.18	–	–	–	–	78.40
Naive Bayes [55]	88.34	–	–	–	–	70.20
MLP [55]	96.78	–	–	–	–	81.24
CNN [55]	95.94	–	–	–	–	82.00
RBM [62]	–	99.31	97.56	90.01	98.13	85.60

Massey: Massey University Gesture Dataset 2012, ASl: American Sign Language (ASL), Surrey: Fingerspelling Dataset from the University of Surrey's Center for Vision, Speech and Signal Processing, NYU, and Fingerspelling A: ASL Fingerspelling A showed that the restricted subject numbers and also the sample numbers per gesture classes can have a negative impact on deep learning-based models. To overcome these constraints, we proposed a dataset including more subject numbers and also more sample numbers for each gesture class to meet deep learning requirements. Different analysis on the exiting state-of-the-art models in HGR, HPE, hand recovery, and hand reconstruction were performed and reported. Overall, we hope this dataset provides a baseline for working in real-world conditions for the research community.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflicts of interests/Competing interests The authors certify that they have no conflict of interest.

References

1. Adarsh P, Rathi P, Kumar M (2020) Yolo v3-tiny: Object detection and recognition using one stage improved model. International Conference on Advanced Computing and Communication Systems (ICACCS)
2. Adithya V, Rajesh R (2020) A deep convolutional neural network approach for static hand gesture recognition. Procedia Comput Sci 171:2353–2361
3. Alani AA, Cosma G, Taherkhani A, McGinnity T (2018) Hand gesture recognition using an adapted convolutional neural network with data augmentation. International Conference on Information Management (ICIM)
4. Ameen S, Vadera S (2016) A convolutional neural network to classify american sign language fingerspelling from depth and colour images. Wiley Expert Systems
5. Baek S, Kim K, Kim TK (2018) Augmented skeleton space transfer for depth-based hand pose estimation. CVPR, Salt Lake City, pp 8330–8339
6. Benitez-Garcia G, Olivares-Mercado J, Sanchez-Perez G, Yanai K (2020) Ipn hand: A video dataset and benchmark for real-time continuous hand gesture recognition. ICPR
7. Blinded (2021) Hand gesture recognition. Accessed Oct 2023. http://shenasa.ai/service/59/hand-gesture-recognition
8. Bloom V, Makris D, Argyriou V (2012) G3d: A gaming action dataset and real time action recognition evaluation framework. Computer Society Conference on Computer Vision and Pattern Recognition Workshops
9. Bochkovskiy A, Wang CY, Liao HYM (2020) Yolov4: Optimal speed and accuracy of object detection. arXiv:2004.10934
10. Cao Z, Simon T, Wei SE, Sheikh Y (2017) Real time multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE conference on computer vision and pattern recognition, Hawaii Convention Center, Honolulu, Hawaii, pp 7291–7299
11. Chen Y, Ma H, Kong D, Yan X, Wu J, Fan W, Xie X (2020) Non-parametric structure regularization machine for 2d hand pose estimation. The IEEE Winter Conference on Applications of Computer Vision (WACV)
12. Chen X, Wanga G, Guoa H, Zhanga C (2108) Pose Guided Structured Region Ensemble Network for Cascaded Hand Pose Estimation. Neurocomputing. https://doi.org/10.1016/j.neucom.2018.06.097
13. Chen Y, Zhao L, Peng X, Yuan J, Metaxas D (2019) Construct Dynamic Graphs for Hand Gesture Recognition via Spatial-Temporal Attention. BMVC, UK pp 1–13
14. Cooper H, Ong EJ, Pugeault N, Bowden R (2012) Sign language recognition using sub-units. J Mach Learn Res 13:2205–2231
15. Cui R, Liu H, Zhang C (2019) A deep neural framework for continuous sign language recognition by iterative training. IEEE Trans Multimed 21(7):1880–1891

16. Dadashzadeh A, Tavakoli Targhi A, Tahmasbi M (2018) HGR-Net: A Two-stage Convolutional Neural Network for Hand Gesture Segmentation and Recognition. arXiv:1806.05653
17. Dibia V (2017) HandTrack: A Library For Prototyping Real-time Hand Tracking Interfaces using Convolutional Neural Networks. GitHub repository. https://github.com/victordibia/handtracking/tree/master/docs/handtrack.pdf
18. dos Santos CC, Samatelo JLA, Vassallo RF (2019) Dynamic gesture recognition by using cnns and star rgb: a temporal information condensation. arXiv:1904.08505v1
19. Duan J, Zhou S, Wany J, Guo X, Li S (2016) Multi-Modality Fusion based on Consensus-Voting and 3D Convolution for Isolated Gesture Recognition. arXiv:1611.06689
20. Elboushaki A, Hannane R, Afdel K, Koutti L (2020) MultiD-CNN: A multi-dimensional feature learning approach based on deep convolutional networks for gesture recognition in RGB-D image sequences. Expert Syst Appl 139:112829
21. Escalera S, González J, Baró X, Reyes M, Guyon IM, Athitsos V, Escalante HJ, Sigal L, Argyros AA, Sminchisescu C, Bowden R, Sclaroff S (2013) Chalearn multi-modal gesture recognition. ICMI '13: Proceedings of the 15th ACM on International conference on multimodal interaction, pp 365–368
22. Cardenas EJE, Chavez GC (2020) Multimodal hand gesture recognition combining temporal and pose information based on CNN descriptors and histogram of cumulative magnitudes. J Vis Commun Image Represent 71:102772
23. Feichtenhofer C, Pinz A, Wildes RP (2016) Spatiotemporal residual networks for video action recognition. NIPS
24. Haque A, Peng B, Luo Z, Alahi A, Yeung S, Fei-Fei L (2016) Towards Viewpoint Invariant 3D Human Pose Estimation. ECCV, Amsterdam
25. Hasson Y, Varol G, Tzionas D, Kalevatykh I, Black M, Laptev I, Schmid C (2019) Learning joint reconstruction of hands and manipulated objects. CVPR
26. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. arXiv:1512.03385
27. Mousavi HH, Khademi M (2014) A review on technical and clinical impact of microsoft kinect on physical therapy and rehabilitation. J Med Eng 846514
28. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H (2017) Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861
29. Huang G, Liu Z, van der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. CVPR
30. John J, Deshpande S (2023) Static hand gesture recognition using multi-dilated DenseNet-based deep learning architecture. Imaging Sci J 71(3): 221–243
31. Kim S, Ban Y, Lee S (2017) Tracking and classification of in-air hand gesture based on thermal guided joint filter. Sensors 17(1):166
32. Koller O, Forster J, Ney H (2013) Continuous sign language recognition: towards large vocabulary statistical recognition systems handling multiple signers. Comp Vision Image Underst 141:108–125
33. Koller O, Ney H, Bowden R (2015) Deep Learning of Mouth Shapes for Sign Language. IEEE International Conference on Computer Vision Workshop (ICCVW), Santiago
34. Kopuklu O, Gunduz A, Kose N, Rigoll G (2019) Real-time hand gesture detection and classification using convolutional neural networks. arXiv:1901.10323
35. Kuniyoshi Y, Inoue H, Inaba M (1990) Design and implementation of a system that generates assembly programs from visual recognition of human action sequences. IEEE International Workshop on Intelligent Robots and Systems, Towards a New Frontier of Applications
36. Kurakin A, Zhang Z, Liu Z (2012) A real time system for dynamic hand gesture recognition with a depth sensor. In 2012 Proceedings of the 20th European signal processing conference (EUSIPCO), Bucharest, Romania, pp. 1975–1979
37. LabelImg (2021) Labelimg: A graphical image annotation tool. Accessed Oct 2023. https://githubcom/tzutalin/labelImg
38. Lang S, Block-Berlitz M, Rojas R (2012) Sign language recognition and translation with kinect. Proceedings of the 11th international conference on Artificial Intelligence and Soft Computing - Volume Part I
39. Li Y, Xue Z, Wang Y, Ge L, Ren Z, Rodriguez J (2019) End-to-End 3D Hand Pose Estimation from Stereo Cameras. BMVC, UK
40. Lim K, Tan A, Lee C, Tan S (2019) Isolated sign language recognition using Convolutional Neural Network hand modelling and Hand Energy Image. Multimedia Tools Appl 78:19917–19944
41. Lin K, Wang L, Liu Z (2021) End-to-end human pose and mesh reconstruction with transformers. CVPR

42. Li J, Su W, Wang Z (2020) Simple pose: Rethinking and improving a bottom-up approach for multi-person pose estimation. The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020)
43. Lv J, Xu W, Yang L, Qian S, Mao C, Lu C (2021) Handtailor: Towards high-precision monocular 3d hand recovery. BMVC
44. Ma1 M, Chen Z, Wu J (2016) A recognition method of hand gesture with cnn-svm model. International Conference on Bio-Inspired Computing: Theories and Applications, Harbin, pp 399–404
45. Majidi N, Kiani K, Rastgoo R (2020) A deep model for super-resolution enhancement from a single image. J AI Data Min 8:451–460
46. Marın-Jimenez MJ, Romero-Ramirez FJ, Munoz-Salinas R, Medina-Carnicer R (2018) 3D human pose estimation from depth maps using a deep combination of poses. J Vis Commun Image Represent 55: 627–639
47. Marks R (2011) System and method for providing a real-time three-dimensional interactive environment. US Patent 8,072,470
48. Mocialov B, Turner G, Lohan K, Hastie H (2017) Towards continuous sign language recognition with deep learning. In Proc. of the Workshop on the Creating Meaning With Robot Assistants: The Gap Left by Smart Devices, 5525834
49. Moghbeli Damaneh M, Mohanna F, Jafari P (2023) Static hand gesture recognition in sign language based on convolutional neural network with feature extraction method using orb descriptor and gabor filter. Expert Syst Appl 211:118559
50. Mohammadi M, Chandarana P, Seekings J, Hendrix S, Zand R (2022) Static hand gesture recognition for american sign language using neuromorphic hardware. Neuromorphic Comput Eng 2(4):044005
51. Mohanty A, Rambhatla S, Sahay R (2017) Deep gesture: Static hand gesture recognition using CNN. Proceedings of International Conference on Computer Vision and Image Processing Advances in Intelligent Systems and Computing
52. Moon G, Chang J, Lee K (2018) V2V-PoseNet: Voxel-to-Voxel Prediction Network for Accurate 3D Hand and Human Pose Estimation from a Single Depth Map. CVPR, Salt Lake City
53. Mueller F, Bernard F, Sotnychenko O, Mehta D, Sridhar S, Casas D, Theobalt C (2018) Ganerated hands for realtime 3d hand tracking from monocular rgb. CVPR, Salt Lake City, pp 1–11. https://doi.org/10.1109/CVPR.2018.00013
54. Newell A, Yang K, Deng J (2016) Stacked Hourglass Networks for Human Pose Estimation. European Conference on Computer Vision (ECCV), pp 483–499
55. Noble F, Xu M, Alam F (2023) Static hand gesture recognition using capacitive sensing and machine learning. Sensors 23(7):3419
56. Oberweger M, Riegler G, Wohlhart P, Lepetit V (2016) Efficiently Creating 3D Training Data for Fine Hand Pose Estimation. CVPR, Nevada
57. Ong EJ, Cooper H, Pugeault N, Bowden R (2012) Sign language recognition using sequential pattern trees. CVPR
58. Oszust M, Wysocki MJ (2013) Polish sign language words recognition with kinect. International Conference on Human System Interaction (HSI'2013)
59. Park J, Oh Y, Moon G, Choi H, Lee K (2022) Handoccnet: Occlusion-robust 3d hand mesh estimation network. CVPR
60. Pugeault N, Bowden R (2011) Spelling it out: Real-time asl fingerspelling recognition. International Conference on Computer Vision Workshops (ICCV Workshops)
61. Rastgoo R, Kiani K, Escalera S (2021) Sign language recognition: A deep survey. Expert Syst Appl 164:113794
62. Rastgoo R, Kiani K, Escalera S (2018) Multi-modal deep hand sign language recognition in still images using restricted Boltzmann machine. Entropy 20(11):809
63. Rastgoo R, Kiani K, Escalera S (2020) Hand sign language recognition using multi-view hand skeleton. Expert Syst Appl 150:113336
64. Rastgoo R, Kiani K, Escalera S (2020) Video-based isolated hand sign language recognition using a deep cascaded model. Multimed Tools Appl 79:22965–22987
65. Rastgoo R, Kiani K, Escalera S (2021) Hand pose aware multimodal isolated sign language recognition. Multimedia Tools Appl 80:127–163
66. Rastgoo R, Kiani K, Escalera S (2022) Real-time isolated hand sign language recognition using deep networks and SVD. J Ambient Intell Humanized Comput 13(1):591–611
67. Rautaray SS, Agrawal A (2012) Real time gesture recognition system for interaction in dynamic environment. Procedia Technol 4:595–599

68. Roccetti M, Marfia G, Zanichelli M (2010) The art and craft of making the tortellino: playing with a digital gesture recognizer for preparing pasta culinary recipes. Comput Entertain 8(4):1–20
69. Roccetti M, Marfia G, Semeraro A (2012) Playing into the wild: A gesture-based interface for gaming in public spaces. J Vis Commun Image Represent 23(3):426–440
70. Romero J, Tzionas D, Black MJ (2022) Embodied hands: Modeling and capturing hands and bodies together. arXiv preprint arXiv:2201.02610
71. Ronchetti F, Quiroga F, Estrebou C, Lanzarini L, Rosete A (2016) Lsa64: A dataset of argentinian sign language. Congreso Argentino de Ciencias de la Computación (CACIC)
72. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) Mobilenetv2: Inverted residuals and linear bottlenecks. CVPR
73. Sharma S, Singh S (2021) Vision-based hand gesture recognition using deep learning for the interpretation of sign language. Expert Syst Appl 182:115657
74. Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. NIPS, pp 1–9
75. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556
76. Spurr A, Song J, Park S, Hilliges O (2018) Cross-modal deep variational hand pose estimation. CVPR, Salt Lake City, pp 89–98
77. Subhashini S, Revathi R (2023) Static and dynamic hand gesture recognition system with deep convolutional levy flight whale optimization. Multimedia Tools Appl. https://doi.org/10.1007/s11042-023-15397-8
78. Szegedy C, Ioffe S, Vanhoucke V, Alemi A (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. AAAI
79. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. CVPR
80. Tan M, Pang R, Le QV (2020) Efficientdet: Scalable and efficient object detection. CVPR
81. Thangali A, Nash JP, Sclaroff S, Neidle C (2011) Exploiting phonological constraints for handshape inference in ASL video. CVPR
82. Hoang VT (2020) HGM-4: A new multi-cameras dataset for hand gesture recognition. Data Brief 30:105676
83. Vishwakarma S, Agrawal A (2013) A survey on activity recognition and behavior understanding in video surveillance. Vis Comput 29:983–1009
84. Wang M, Chen X, Liu W, Qian C, Lin L, Ma L (2018) DRPose3D: Depth Ranking in 3D Human Pose Estimation. Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), pp 978–984
85. Wang J, Liu Z, Wu Y, Yuan J (2012) Mining actionlet ensemble for action recognition with depth cameras. CVPR
86. Wang L, Qiao Y, Tang X (2015) Action recognition with trajectory-pooled deep-convolutional descriptors. CVPR, pp 4305–4314
87. Wei S, Ramakrishna V, Kanade T, Sheikh Y (2016) Convolutional Pose Machines. CVPR, Las Vegas
88. with Depthwise Separable Convolutions XDL (2017) François chollet. arXiv:1610.02357
89. Yamato J, Ohya J, Ishii K (1992) Recognizing human action in time-sequential images using hidden markov model. Proceedings 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition
90. Yu J, Qin M, Zhou S (2022) Dynamic gesture recognition based on 2d convolutional neural network and feature fusion. Sci Rep 12(1):4345
91. Zimmermann C, Brox T (2017) Learning to estimate 3d hand pose from single RGB images. ICCV
92. Zoph B, Vasudevan V, Shlens J, Le QV (2018) Learning transferable architectures for scalable image recognition. CVPR
93. Rastgoo R, Kiani K, Escalera S (2023) ZS-GR: zero-shot gesture recognition from RGB-D videos. Multimed Tools Appl. https://doi.org/10.1007/s11042-023-15112-7
94. Rastgoo R, Kiani K, Escalera S (2022) Word separation in continuous sign language using isolated signs and post-processing. arXiv:2204.00923
95. Rastgoo R, Kiani K, Escalera S, Sabokrou M (2022) Multi-modal zero-shot sign language recognition. arXiv:2109.00796
96. Rastgoo R, Kiani K, Escalera S, Sabokrou M (2021) Sign language production: A review, Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3451–3461

97. Rastgoo R, Kiani K, Escalera S, Athitsos V, Sabokrou M (2022). All You Need In Sign Language Production. arXiv:2201.01609