Multimedia Tools and Applications (2020) 79:22965–22987
https://doi.org/10.1007/s11042-020-09048-5

Video-based isolated hand sign language recognition using a deep cascaded model

Razieh Rastgoo[1] · Kourosh Kiani[1] · Sergio Escalera[2]

Received: 26 June 2019 / Revised: 9 March 2020 / Accepted: 7 May 2020 /
Published online: 2 June 2020

Abstract
In this paper, we propose an efficient cascaded model for sign language recognition taking benefit from spatio-temporal hand-based information using deep learning approaches, especially Single Shot Detector (SSD), Convolutional Neural Network (CNN), and Long Short Term Memory (LSTM), from videos. Our simple yet efficient and accurate model includes two main parts: hand detection and sign recognition. Three types of spatial features, including hand features, Extra Spatial Hand Relation (ESHR) features, and Hand Pose (HP) features, have been fused in the model to feed to LSTM for temporal features extraction. We train SSD model for hand detection using some videos collected from five online sign dictionaries. Our model is evaluated on our proposed dataset (Rastgoo et al., Expert Syst Appl 150: 113336, 2020), including 10’000 sign videos for 100 Persian sign using 10 contributors in 10 different backgrounds, and isoGD dataset. Using the 5-fold cross-validation method, our model outperforms state-of-the-art alternatives in sign language recognition

Keywords Sign language · Deep learning · Dataset · Single Shot Detector (SSD) · Hand pose · Video

1 Introduction

Nowadays that the communication technologies and tools such as Imo [24] and WhatsApp [1] have become an important part of our life, they can be used to facilitate the

✉ Kourosh Kiani
Kourosh.kiani@semnan.ac.ir

Razieh Rastgoo
rrastgoo@semnan.ac.ir

Sergio Escalera
sergio@maia.ub.es

[1] Electrical and Computer Engineering Department, Semnan University, Semnan, Iran

[2] Department of Mathematics and Informatics, University of Barcelona and Computer Vision Center, UAB, Barcelona, Spain

communication between deaf community and hearing majority. While deaf people could communicate with each other using these technologies, they have many problems for communicating with people who do not know sign language. So, development of automatic sign language translation systems is necessary to provide equal communication opportunity and improve public welfare. With the advent of deep learning in recent years, many research efforts have been conducted to sign language recognition [8, 12, 20, 26, 40, 41]. There are still some challenges in this area such as high occlusions of hands, fast hands movement, background complexity, inexistence of the large and diverse datasets, varying illumination conditions, different hand gestures, and complex interactions between hands and objects. In this paper, we propose a deep model for efficient hand sign recognition including the following contributions:

- **Dataset**: With the advent of deep learning in recent years, many deep-based models have been proposed in sign language recognition area. These models need to be learned using large and diverse datasets in order to benefit from deep learning capabilities. There are many datasets, with different data modalities and languages, for hand sign language recognition. While deep-based models need datasets including large numbers of data samples in each sign label, the current datasets do not meet this requirement. In this paper, we use our proposed dataset [27] including more sample numbers in each sign label in comparison with the current hand sign datasets. Our dataset contains 10’000 RGB videos for 100 Persian sign words in 10 different environments using 10 contributors for signing.
- **Complexity and efficiency**: There are some accurate models for hand sign language recognition that suffer from high model complexity. In this paper, we propose a simple yet efficient and accurate cascaded model including Single Shot Detector (SSD), Convolutional Neural Network (CNN) and Long Short Term Memory (LSTM) for hand detection and sign recognition from video. To provide convenient and easy communication between deaf and hearing communities in real-time applications, we need to have not only an accurate but also fast enough model to allow for real-time communication.
- **Accuracy**: Many deep models have been provided for sign language recognition. In this paper, we propose a deep cascaded model for robust hand detection and sign recognition with higher performance than the current state-of-the-art models in a more naturalistic environment. We train SSD model for hand detection using a large amount of annotated data. We achieve state-of-the-art results on the proposed dataset and isoGD.

The remainder of this paper is organized as follows. Section 2 presents recent related work in sign language recognition area. Details of the proposed model are explained in Section 3. We report the evaluation results in Section 4. Finally, we conclude the work in Section 5.

2 Related work

While hand sign recognition models have been rapidly advanced in recent years, there are some challenges that need to be solved. We review some related works for sign language recognition from three perspectives as follows:

- **Dataset**: There are different hand sign language datasets, including different human body parts, with different details of input modality, language, sample numbers, class numbers, environmental and spatial constraints, and contributor numbers. In this paper, we use our proposed dataset including 10’000 RGB videos for 100 Persian sign words

Table 1 Sign language datasets for video inputs. 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

Y	Dataset	C	CN	SubN	SampN	SPC	LL	A	Availability
2011	Boston ASL LVD [35]	USA	3300	6	9800	3	W	H	Public
2012	DGS Kinect 40 [4]	Germany	40	15	3000	8	W	–	Public
2012	RWTH-PHOENIX-Weather [7]	Germany	1200	7	45760	5	S	F, H	Public
2012	GSL 20 [22]	Greek	20	6	840	5	W	–	Public
2013	PSL Kinect 30 [23]	Poland	30	1	300	10	W	–	Public
2013	PSL ToF 84 [23]	Poland	84	1	1680	10	W	–	Public
2014	DEVISIGN-G [2]	China	36	8	432	5	W	–	Public
2014	DEVISIGN-D [2]	China	500	8	6000	5	W	–	Public
2014	DEVISIGN-L [2]	China	239	8	24000	5	W	–	Public
2015	SIGNUM [15]	Germany	455	25	33210	3	S	–	Public
2016	MSR [3]	USA	12	10	336	3	W	–	Public
2016	LSA64 [30]	Argentina	64	10	3200	5	W	H, h	Public
2016	TVC-hand gesture [14]	Korea	10	1	650	5	–	–	Public
2020	RKS-PERSIANSIGN [27]	Iran	100	10	10’000	100	W	H	Will be Public

in 10 different environments using 10 contributors for signing. We have 100 samples for each sign words. Details of the sign language datasets including our dataset have been shown in Tables 1 and 2. As one can see in Table 1, while four datasets [2, 7, 15, 35] include high sample numbers, 9800, 45760, 24000, 33210, they have many class numbers, 3300, 1200, 2000, 450, that led to have only 3, 5, 5, 3 samples for each sign. We increased the sample numbers to 100 for each sign in order to improve the recognition accuracy of deep-based models.

- **Complexity and efficiency**: In this category, we review some recent deep models from complexity perspective. Ge et al., present a multi-view CNN-based model to project the depth image into three orthogonal planes and regress the 2D heat-maps of them in order to estimate the hand joint positions in real-time. The final 3D hand pose with learned pose priors is estimated using multi-view heat-maps. While the evaluation results on a dataset of [34] show that the proposed method is fast and outperforms state-of-the-art models in real-time hand pose estimation area [8], this model suffers from high parameters complexity. Zhou et al., propose a CNN-model by features fusion of three parts of a hand including thumb, index finger and the other fingers to present differences in the functional importance of different fingers. A feature ensemble layer along with a low-dimensional embedding layer has been used to apply the overall hand shape

Table 2 Sign language datasets for input images

Year	Dataset	Con	CN	Sub	Samp.
2011	ASL Fingerspelling A [25]	USA	24	5	131000
2011	ASL Fingerspelling B [25]	USA	24	9	–
2016	LSA16 handshapes [29]	Argentina	16	10	800
2015	PSL Fingerspelling ToF [13]	Poland	16	3	960

constraints to the model. They evaluated the model on HIM2017, ICVL, and MSRA datasets and achieved the comparable performance to state-of-the-art methods with less parameter complexity and faster frame rate [40]. Rastgoo et al., have proposed a hand sign recognition model using RBM from two modalities. They use three forms of input data: original image, cropped image, and noisy cropped image. In the first step, the hand of each crop is detected using a CNN. After that, for each modality, three forms of an input image are input to RBMs. The outputs of the RBMs for two modalities are fused in another RBM in order to recognize the output sign label. The proposed multimodal model is trained on four publicly available datasets, Massey University Gesture Dataset 2012, Fingerspelling Dataset from the University of Surrey's Center for Vision, Speech and Signal Processing, NYU, and ASL Fingerspelling A. While the evaluation results showed that the model has achieved the state-of-the-art results for recognition accuracy from still images, they need to decrease the complexity of the model by sharing the parameters [26]. As one can see in the third column of Table 3, while the reviewed fast models in this category are considered as the state-of-the-art models in hand sign recognition area, they still suffer from high model complexity. To decrease the model complexity with preserving or improving the model accuracy, we propose a simple yet accurate deep-based cascaded model using SSD, CNN, and LSTM. We train SSD model for hand detection using five online sign dictionaries to improve the detection accuracy. To best of our knowledge, this is the first time that SSD model is trained for sign language area. Furthermore, we fused three feature types, hand features extracted by CNN, ESHR, and HP to feed them to LSTM for temporal feature extraction. Our model is not only simple and accurate but also has lower parameters complexity in comparison with state-of-the-art models. As the third column of Table 3 shows, We decrease the model complexity, including the complexities of different parts of the proposed model, in comparison with the other models.

- **Accuracy**: In this category, we review deep sign language recognition models from recognition accuracy perspective. Zimmermann and Brox provide a CNN-based model to learn 3D articulation priors and keypoints from an RGB input image. Their model contains three main steps for localization, cropping, and estimation of the hand that CNN is applied in all steps. In addition, they propose a large scale 3D hand pose dataset based on the synthetic hand models [41]. While the evaluation results of the model on two datasets, Stereo Hand Pose Tracking Benchmark and Dexter, show that the model performance is even competitive to state-of-the-art models for hand pose estimation, their model needs to be learned by real-world images and diverse pose statistics. Kang et al., present a CNN-based model for real-time sign language recognition from a single depth map. They use different learning configurations for model training and evaluate the model on some recorded depth videos. They report state-of-the-art results in recognition accuracy on own dataset [12]. While the model is accurate for observed signers

Table 3 Model complexity of the recent works

Ref.	Model	Complexity	Year
[8]	3D CNN	1.592 GB	2018
[40]	CNN	67.27 MB	2018
[26]	RBM	62.5 MB	2018
Ours	SSD, CNN, LSTM	49.7 MB	2019

Table 4 Model accuracy of the recent works

Ref.	Model	Accuracy	Year
[41]	CNN	66.8	2017
[12]	CNN	85.49	2015
[20]	CNN	82.07	2018
Ours	SSD, CNN, LSTM	98.42	2019

in real-time, it needs to be trained with more data from different subjects to improve the results. Narayana et al., provide a multi-channel CNN-based model in which different channels process different data modalities. Three spatial attention regions, one for the whole input image and one for each of the hands, have been dedicated to these channels. They use the combinations of body and hand movements to reflect the gestures structure. While the evaluation results of the model on isoGD dataset show that this model achieves state-of-the-art results for gesture recognition, it does not address the temporal information [20]. So in this way, we propose a deep cascaded model, including SSD, CNN, and LSTM, for hand sign recognition from video. The evaluation results on our dataset and isoGD show the accuracy improvement of our model in comparison with state-of-the-art models in hand sign recognition area, as one can see in third column of Table 4.

In the next section, we explain the details of our model.

3 Proposed model

In this work, we use the combinational capabilities of three deep models, SSD, CNN, and LSTM, for hand sign recognition. As Fig. 1 shows, the proposed cascaded model includes two main steps as follows:

- **Hand detection:** In this step, we train SSD model for hand detection using five online sign dictionaries videos.
- **Sign recognition:** Three spatial features of detected hands from previous step are fused in this part of the model to feed to a LSTM for temporal feature extraction and sign recognition.

In the following subsections, we explain the details of these steps in the proposed model.

3.1 Hand detection

Hand detection, as the first step of hand sign language recognition, currently has a significant role in this area. However, many researches have been conducted to improve hand detection models [21, 32, 38], this task still includes many challenges in computation time and detection accuracy aspects [16]. Some of the well-known object detection methods have been fine-tuned using transfer learning approach in order to use their strong capabilities in hand detection area [17, 28]. Here, we use SSD [17], as an accurate and fast model for accurate hand detection. SSD is a feed-forward convolutional network that predicts the objects bounding boxes along with the classes scores using small convolutional filters applied to the feature maps. A Non-Maximum Suppression (NMS) step is used in the final step to estimate the final detection. To improve the detection accuracy of SSD model, We

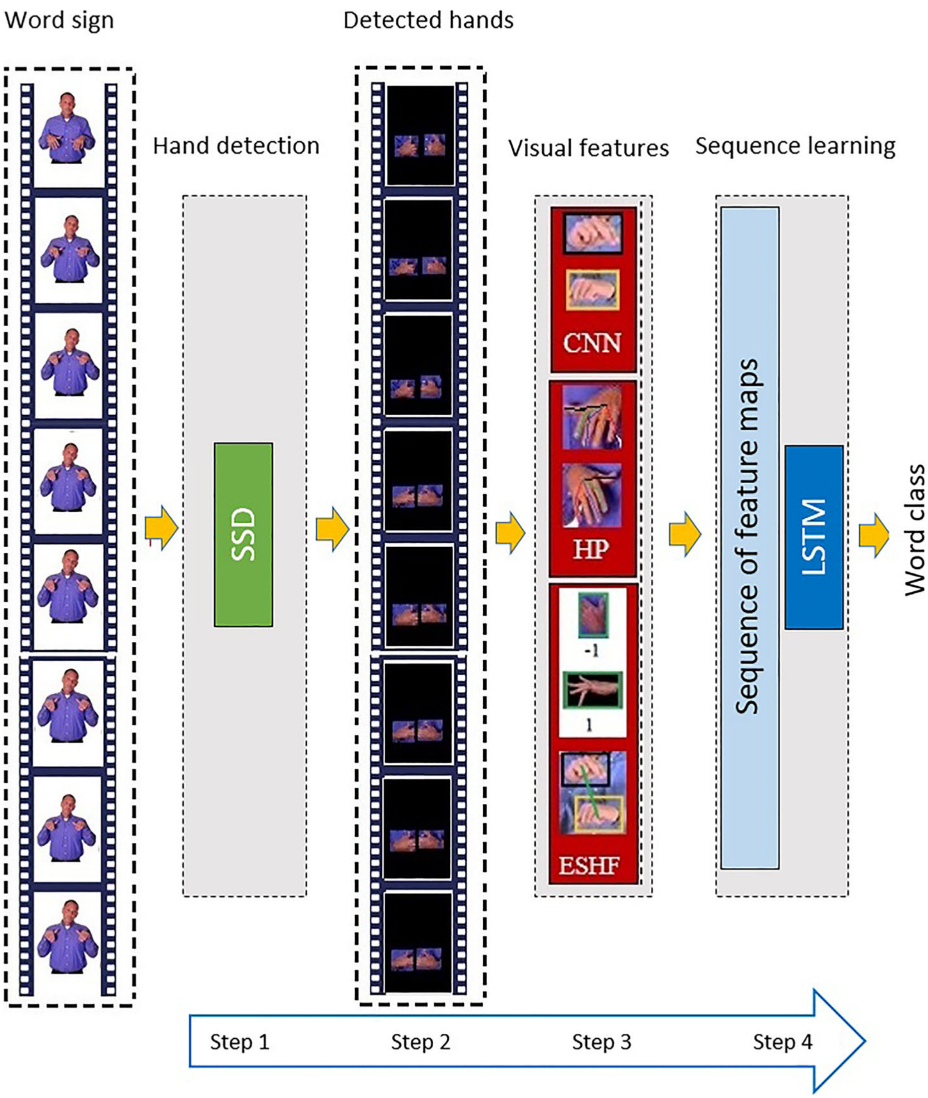


Fig. 1 The proposed model for hand sign language recognition from input video

provide a large and annotated video dataset using five online sign dictionaries [9, 11, 19, 31, 37] to train SSD model.

3.2 Sign recognition

We propose an efficient model using cascaded architecture of SSD, CNN, and LSTM taking benefit from spatio-temporal hand-based information. As we commented, our model includes two main steps: hand detection and sign recognition. We use the detected hands, discussed in the previous subsection, for hand sign recognition in this step. Due to the significant capabilities of CNN for features extraction from still images, we use a CNN, especially ResNet50 model [10], to provide the high-level features of each input frame. We fuse these features with ESHR and HP features to feed them to a LSTM for temporal feature extraction. Details of different parts of this step are explained in the following sub-sections.

3.2.1 Spatial feature extraction

To benefit from the relation features between two hands during signing, we use three types of spatial features in our model as follows:

- **Hand features**: After hand detection using SSD, we use ResNet50 model for hand feature extraction. We did not use the Fully Connected layer of the ResNet50 model in our model.
- **ESHR features**: To benefit from the features of interaction between two hands during signing, we define some spatial features as follows:
    - **Slope**: Most of the video frames include two hands during signing. Slope feature, as a spatial feature related to interaction of two hands, is defined as the shape of the line connecting two hand boxes centers. Figure 2 shows this feature. If we consider two hand boxes centers as C1 and C2 with (x1,y1) and (x2,y2) coordinates, the Slope feature is calculated as follows:

$$Slope = |y2 - y1| / |x2 - x1| \tag{1}$$

Since the adjacent frames include roughly similar information, if we have only one hand in the input frame, we consider the box of the other hand from previous frame and calculate the Slope feature in this way. When we confront with

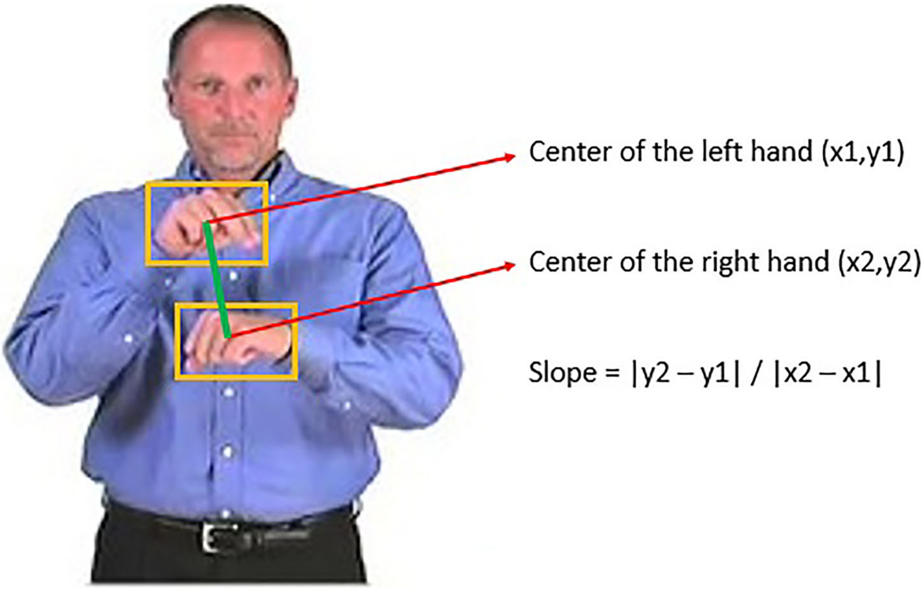

Center of the left hand (x1,y1)
Center of the right hand (x2,y2)
Slope = |y2 – y1| / |x2 – x1|

Fig. 2 The first feature, Slope, considered as extra spatial hand relation feature

an undefined value for slope feature, due to having a same x coordinates for two hands, we use the value of the slope feature in the previous frame. In the case of the first frame, we use a default value for this feature.

- **Orientation**: To use the hand direction features during signing, Orientation feature is defined for detected hand boxes. comparing the width and height of a hand box with each other, we consider two values to assign to the Orientation feature: 1 and -1. We have the following cases:
  - **Case 1**: In this case, we have a horizontal box. So, the width value of the hand box is more than the height one. We assign the value of 1 to the Orientation feature.
  - **Case 2**: In this case, we have a vertical box. So, the height value of the hand box is more than the width one. We assign the value of -1 to the Orientation feature.
  - **Case 3**: In this case, we have a square box. So, the width value of the hand box is equal to the height one. We assign the value of 1 to the Orientation feature.

  Figure 3 shows some samples of this feature.

- **HP features**: To boost the fused features of our model and improve the recognition accuracy, we use hand pose features of the input videos. For this goal, we use the accurate model proposed by Zimmermann and Brox [41] including a CNN-based model for 3D hand pose estimation from an RGB input. They use a CNN-based hand segmentation and localization model to predict the 3D hand keypoints from 2D images. Their model includes three parts for this prediction: hand segmentation and localization, 2D keypoints prediction, 3D keypoints prediction.

3.2.2 Feature fusion

We fuse the extracted features of model including hand features, ESHR features, and HP features to feed them to LSTM for temporal feature extraction. One can see this step in third part of our model in Fig. 1.

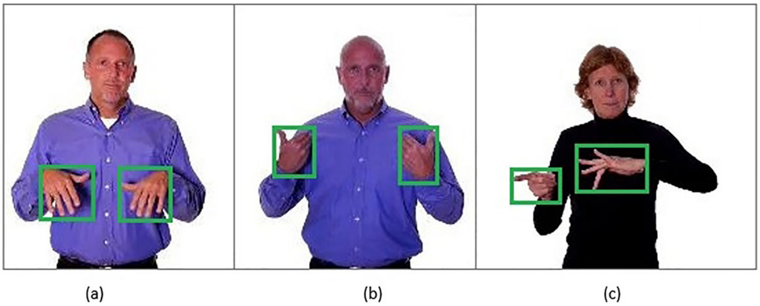


Fig. 3 Second feature, Orientation, considered as extra spatial hand relation feature. **a**: Orientation feature values: 1, 1, **b**: Orientation feature values: -1, -1, **c**: Orientation feature values: 1, 1

3.2.3 Temporal feature extraction

All of the features used in the model, including hand features, ESHR features, and HP features, are fused and flattened to feed to LSTM. These fused features have the length of 713, including the ResNet50 features with 512 length, ESHR features with 75 length, and HP features with 126 length. We use LSTM with experimentally 128 hidden units, as fourth part of the Fig. 1 shows. To make a consistency between the dimension of three spatial features, we repeat each of the values of three ESHR features 25 times, as an experimental value, to have a feature vector with 75 dimensions. Due to have a 3D estimation of hand keypoints, so we have a feature vector including 126 coordinates of 21 keypoints for each video frame.

3.2.4 Final hand sign recognition

We use a Fully Connected (FC) layer with 100 units after LSTM for word sign classification. We use a Softmax layer for 100 Persian words recognition in our model.

3.2.5 Sampling

Since the input videos include different frame numbers, we need to handle these variations by selecting just the predefined frame numbers in each video. There are some methods for this goal such as truncation, summarizing, and sampling [6]. Here, we use sampling method in our model. We calculate the Average Frame Numbers (AFN) of input videos and consider it as the predefined length for input videos. AFN is computed using All Frame Counts (AFC) and Video Numbers (VN), as the following equation:

$$AFN = AFC/VN. \tag{2}$$

So, we select the video frames starting from the first to AFN. Figure 4 shows our sampling method. In the case of having fewer frame numbers than the AFN value, we simply repeat the last frame in order to have the frame numbers equal to AFN value. Also, the suitable value for Frame Per Second (FPS) is very important and depends on the application or project. In other words, we have to manipulate it in order to fit it on our project frame rate.

4 Proposed dataset

To benefit from deep learning approaches, we need to have a large dataset including large amount of samples per each class. The current datasets do not meet this requirement. So, We propose a dataset, including 10’000 videos of 100 Persian signs using 10 contributors in 10 different backgrounds with maximum distance of 1.5 meter between the contributor and camera. Our distance assumptions in the proposed dataset have been considered for the applications suitable for this distance such as sign translation on mobile device, laptop, and other electronic devices. Different adjustments are essential for different applications with different distance assumptions. We use this dataset to train and evaluate of our model for hand sign language recognition from video. We have 100 video samples for each Persian word sign. Our deaf contributors include 5 women and 5 men. We tried to use 100 words of the most usable words in daily communication of deaf people in our sign labels. Some samples of this dataset can be found in the second row of Fig. 5.

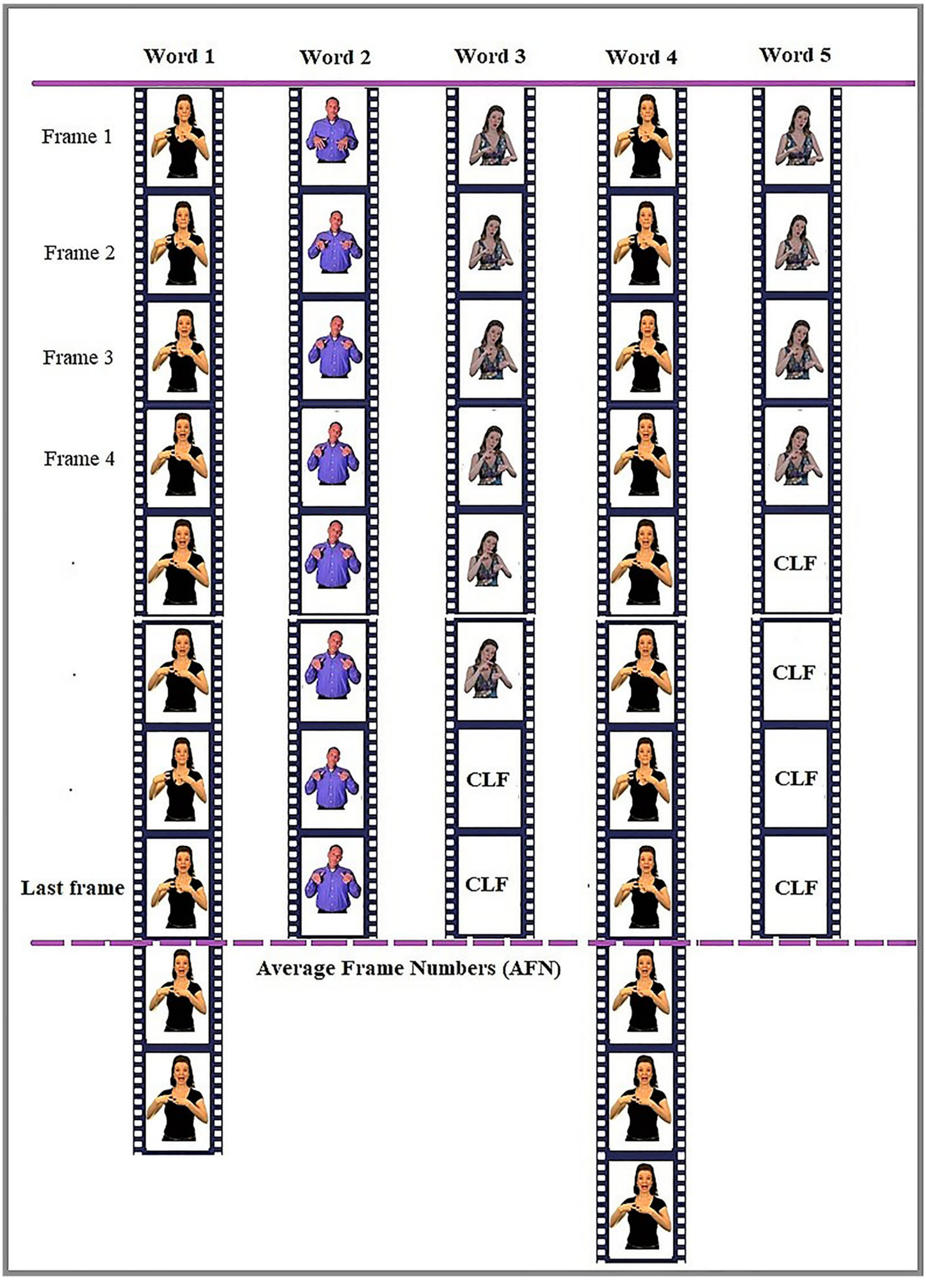


Fig. 4 Our sampling method. We Copy the Last Frame (CLF) in the case of having fewer frame numbers than AFN

5 Results

In this section, we present the experimental results of the proposed model in different steps.

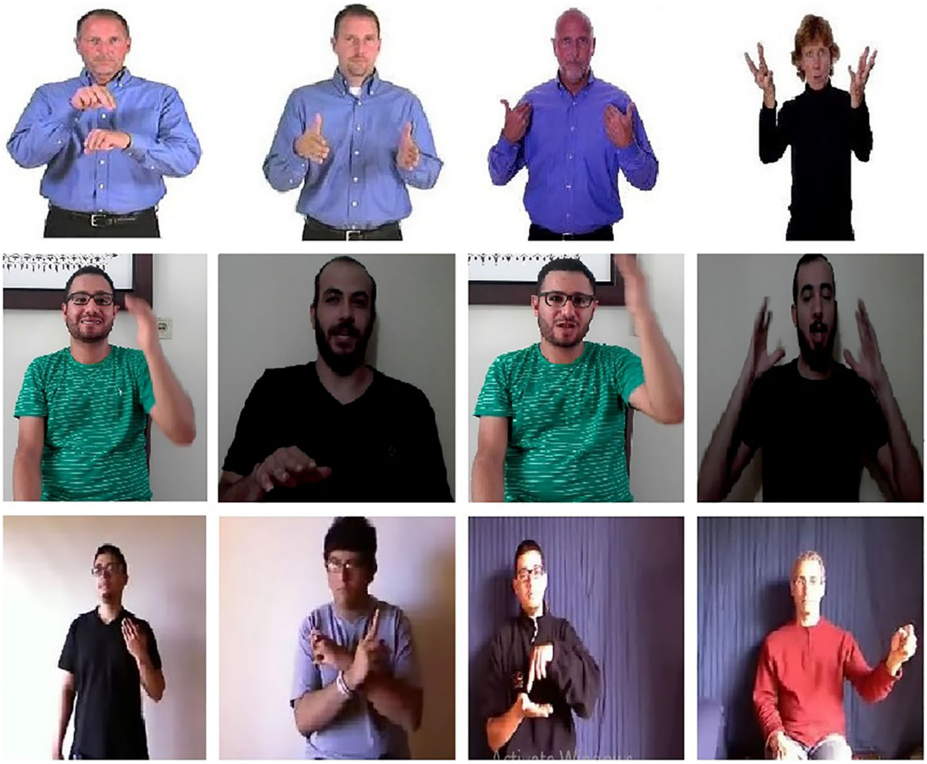


Fig. 5 First row: Some samples of frame videos used for SSD training, Second row: Some samples of our dataset, Third row: Some samples of isoGD dataset

5.1 Implementation details

Our evaluations have been done on Intel(R) Xeon(R) CPU E5-2699 (2 processors) with 30GB RAM on Microsoft Windows 10 operating system and Python software on NVIDIA GPU. We implemented our model on Keras library. We use Stochastic Gradient Descent (SGD) with a mini-batch size of 128. The learning rate starts from 0.005 and is divided by 10 every 1000 epochs. The proposed model is trained for total 10000 and 20000 epochs. In addition, we use a weight decay of 1e-4 and a momentum of 0.92. Table 5 shows the details of the model parameters experimentally set.

Table 5 Details of the parameters used in the proposed model

Parameter	Value	Parameter	Value
Weight decay	1e-4	Iteration numbers	10000, 20000
Learning rate	0.005	Average Frame Numbers (AFN)	50
Batch size	128	Processing way	GPU

Table 6 Details of five dictionaries

Dictionary	Level	Lang.	Link
ASLU [37]	Word	English	http://www.lifeprint.com/index.htm
Baby Sign [9]	Word	English	https://www.babysignlanguage.com/dictionary/?v=04c19fa1e772
Signing Savvy [19]	Word	English	https://www.signingsavvy.com/
Texas Math [31]	Word	English	http://www.tsdvideo.org/
ASLPro [11]	Word	English	http://www.aslpro.com/

5.2 Dataset

We use three datasets for model training and evaluation. Details of our dataset can be found in the previous section. Here, we explain the details of the datasets used for SSD training and model evaluation.

5.2.1 Dataset for SSD training

Due to less availability of a large annotated hand sign video dataset, we annotated the sign videos of five online dictionaries and used them for SSD training to improve the hand detection accuracy. In this step, we only need to improve the hand detection accuracy and the sign label is not important for us. Details of these dictionaries and some samples of the video frames have been shown in Table 6 and first row of Fig. 5.

In the final dataset, we have 106’518 frames with their annotations to train the SSD model.

5.2.2 isoGD dataset

Since we do not have a hand sign dataset containing the similar assumptions to our dataset, we use a gesture recognition dataset including roughly similar assumptions to our dataset in

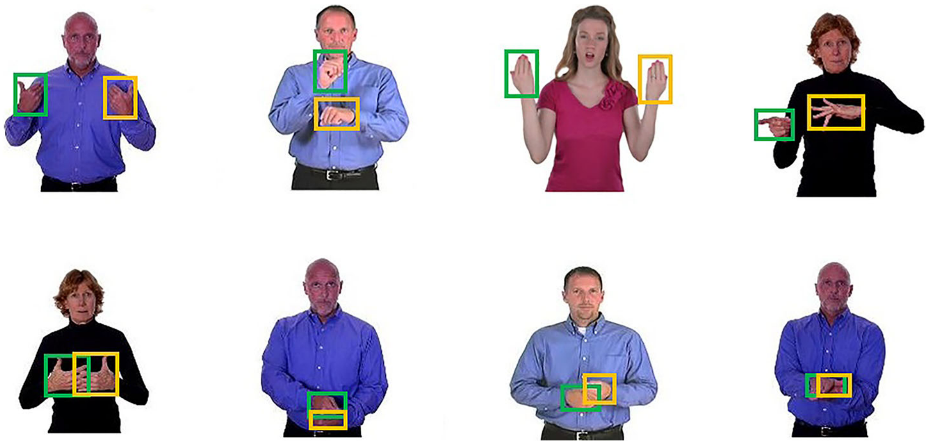


Fig. 6 First row: Some samples of the detected hands using the trained SSD model, Second row: Some samples of input video frames with high occlusion

order to evaluate our model. isoGD dataset includes 47933 RGB and Depth video samples in 249 class labels that we only use the RGB samples. This dataset has been divided to three sub-datasets, including 35878 samples for training, 5784 samples for validating, and 6271 samples for testing. Third row of the Fig. 5 shows some samples of this dataset.

5.3 Hand detection results

We trained and used SSD model for hand detection in our model. Some samples of the detected hands using the trained SSD model can be found in first row of Fig. 6. In the case of having high occlusion in hands, the trained SSD model is facing up with problem that it decreases the detection accuracy. We think that we could more improve the detection accuracy for occluded hands using more annotated data with different distance assumptions and occlusions. Some frame samples with high occlusion are shown in second row of Fig. 6.

5.4 Hand pose estimation results

We use hand pose features in our model to improve the spatial features of the model. As we explained the details of these features in the previous section, we use the accurate model proposed by Zimmermann and Brox [41] for hand pose features extraction. Some samples of the results for this model on our dataset have been shown in Fig. 7. As one can see in this figure, the hand pose features have been accurately estimated in the frame samples including one or two hands.

5.5 Sign recognition results

In this step, a combinational model is used for sign recognition from detected hands of the previous step. A complete analysis has been done to use different spatial features, pre-train models, and temporal-based models, in this part of the model, as follows:

- **Spatial features:** We tried to use three types of spatial features in our model in an incremental approach. In the first step, we only used the CNN hand features. After

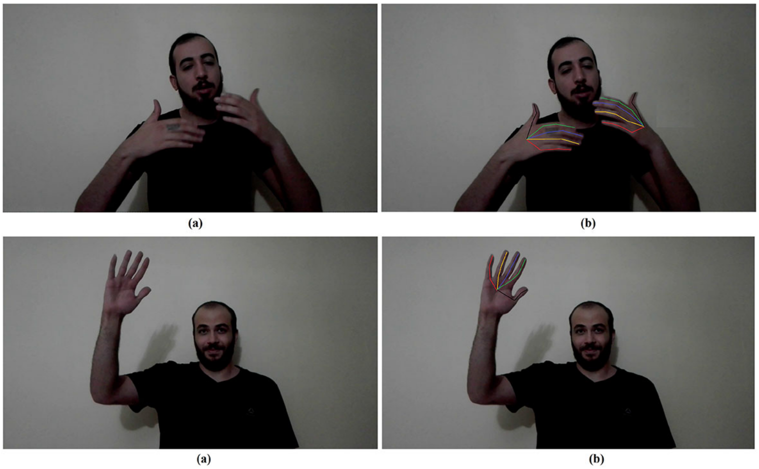


Fig. 7 Some samples of hand pose estimation on our dataset. **a**: Original frame, **b**: Estimated hand pose

that, we included some spatial hand relation features to the model. In the final step, hand pose features have been included in the model to boost the spatial features of our model.

- **Pre-train models:** To select a pre-train CNN model, we analyzed the existence pre-trained models in order to use one of them in our model. We used VGG16 [33] and ResNet50 in our model and analyzed the results. ResNet50 is more accurate with less parameters complexity than VGG16. After trying to use VGG16 model as a pre-trained CNN, we substituted it with ResNet50 in our model.
- **Temporal-based models:** We analyzed LSTM and GRU to take one of them in our model for temporal feature learning. While LSTM is more accurate, GRU is faster due to have less parameters. We use a LSTM in our final model to boost the recognition accuracy of our model.

Our incremental approach, as Fig. 8 shows, to select the final model are as follows:

- SSD+VGG16+LSTM: The model including only hand features extracted by VGG16 fed to LSTM.
- SSD+ResNet50+LSTM: The model including only hand features extracted by ResNet50 fed to LSTM.
- SSD+VGG16+ESHR+LSTM: The model including hand features extracted by VGG16 fused with some extra spatial features related to hands interaction, fed to LSTM.
- SSD+ResNet50+ESHR+LSTM: The model including hand features extracted by ResNet50 fused with some extra spatial features related to hands interaction, fed to LSTM.
- SSD+VGG16+ESHR+HP+LSTM: The model including hand features extracted by VGG16 fused with some extra spatial features related to hands interaction, and hand pose features fed to LSTM.

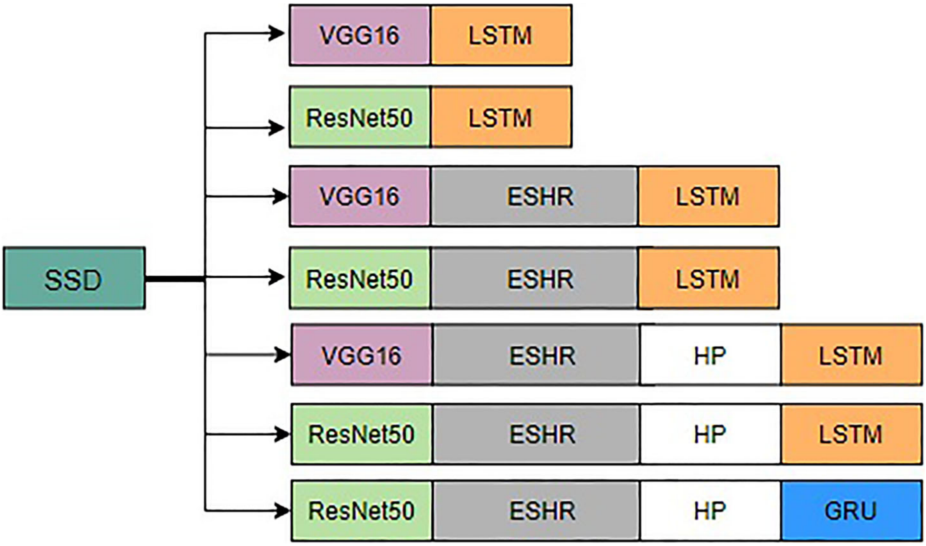


Fig. 8 All of the proposed models for hand sign recognition

- SSD+ResNet50+ESHR+HP+LSTM: The model including hand features extracted by ResNet50 model, fused with some extra spatial features related to hands interaction, and hand pose features fed to LSTM.
- SSD+ResNet50+ESHR+HP+GRU: The model including hand features extracted by ResNet50 fused with some extra spatial features related to hands interaction, and hand pose features fed to GRU.

Our final model includes SSD, ResNet50, and LSTM with three spatial features of CNN hand features, ESHR, and HP. We use two approaches for model training as follows:

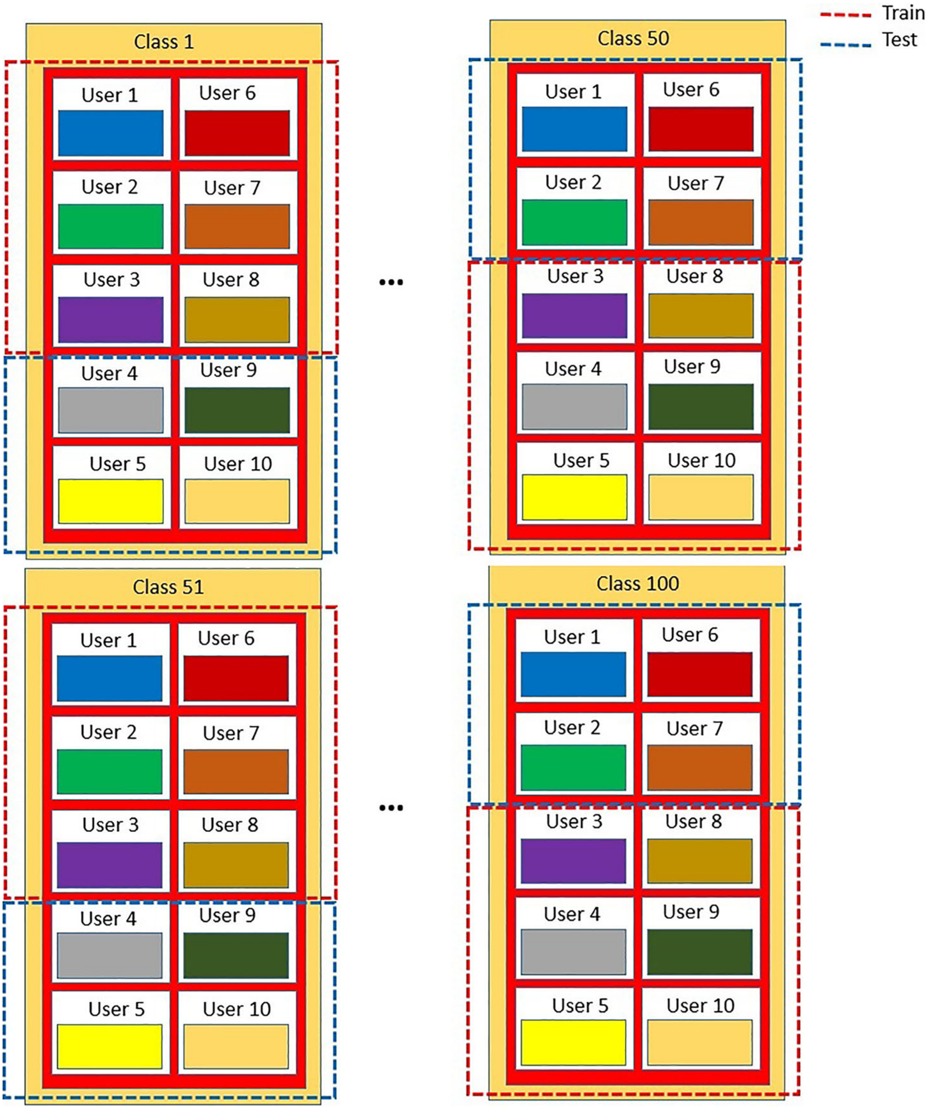


Fig. 9 Visualization of first approach used for data splitting

- **User independent:** In this approach, the contributors in training set will not disappear in testing set. we randomly use 80 percentage of video samples in each sign for training set and 20 percentage for testing set.
- **User dependent with 5-fold cross validation method:** In this approach, we use 5-fold cross validation to train and test the model. We evaluate the model 5 times that in each time we consider all video samples of 8 contributors for train and 2 contributors for test. So, the signers of train and test data are different from each other.

We used both previous approaches for model training. In the first approach, we randomly used 80 percentage of video samples in each sign for training set and 20 percentage for testing set without paying attention to the users. In the second approach, we evaluated the model 5 times. At each time we considered all video samples of 8 contributors for train and 2 contributors for test. So, the signers of train and test data were different from each other. Indeed, while in the first approach we divide all videos of each sign into training and testing sets without caring about the signers, the second approach divides all videos into training and testing sets based on the signers. In other words, we include all videos of 8 signers into training set and 2 signers for testing set. So, in this approach, the proposed model will face with unseen signers in test phase. Figures 9 and 10 show the visualization of these approaches.

Also, we used the average frame numbers, according to Eq. 1, for frame sampling in our experimental results. For 50000 frames of 10000 video samples, we will have the value of 50 for AFN parameter.

5.6 Self-comparison

We train and evaluate the proposed models on our dataset and isoGD using two approaches with two different iteration numbers and report the results in Tables 7 and 8. To have a fair comparison with the other models, we used the isoGD dataset for training and testing in a way of similar to the other works. As one can see in these tables, the model including SSD, ResNet50, and LSTM with the fused features of CNN for hand, ESHR, and HP, has the highest recognition accuracy on both datasets.

5.7 Comparison with state-of-the-art models

We compare the complexity and accuracy of our model with state-of-the-art models in sign language recognition in the following sub-sections.

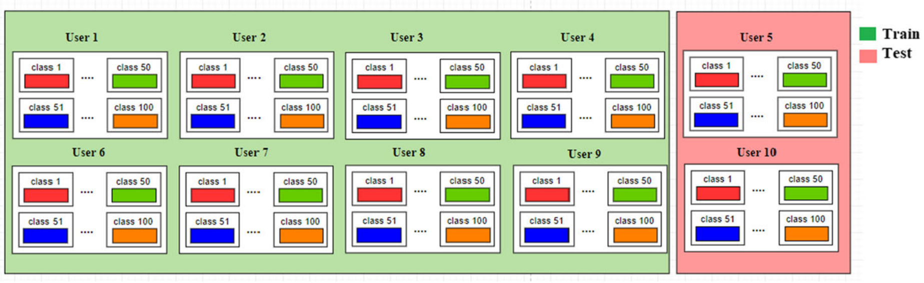


Fig. 10 Visualization of second approach used for data splitting

Table 7 Results of the proposed models on our dataset

Model	Iteration	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Average	All data
SSD+VGG16+LSTM	10000	95.51	95.71	95.62	95.84	95.79	95.69	95.42
SSD+VGG16+LSTM	20000	96.30	96.35	96.25	96.40	96.30	96.32	96.00
SSD+ResNet50+LSTM	10000	96.76	96.54	96.45	96.50	96.65	96.58	96.31
SSD+ResNet50+LSTM	20000	96.98	96.86	96.90	96.82	96.78	96.86	96.82
SSD+VGG16+ESHR+LSTM	10000	96.91	96.86	96.80	96.84	96.86	96.85	96.75
SSD+VGG16+ESHR+LSTM	20000	96.98	96.96	96.92	96.94	96.88	96.93	96.85
SSD+ResNet50+ESHR+LSTM	10000	97.13	97.04	97.08	97.10	97.00	97.07	97.01
SSD+ResNet50+ESHR+LSTM	20000	97.23	97.15	97.18	97.20	97.25	97.20	97.15
SSD+ResNet50+ESHR+HP+LSTM	10000	98.05	98.02	98.08	98.00	98.02	98.03	98.00
SSD+ResNet50+ESHR+HP+LSTM	20000	**98.42**	98.36	98.30	98.38	98.40	98.37	98.28
SSD+ResNet50+ESHR+HP+GRU	20000	97.94	97.86	97.98	98.00	97.92	97.94	97.54

5.7.1 Computational complexity

Our model includes 49.7 MB parameters complexity. We compare the complexity of our model with state-of-the-art models for hand pose estimation. As Table 4 shows, our model has a lower complexity than state-of-the-art models.

5.7.2 Accuracy

We compare the model accuracy in two parts of our model as follows:

- **Hand detection accuracy**: Since we trained SSD model on five online sign dictionaries, we compare the hand detection accuracy of SSD before and after the training to analyze the results. We use the Intersection Over Union (IOU) metric to measure the detection accuracy of SSD before and after the training. IOU measures the overlap between predicted hand box and ground-truth one. As one can see in Fig. 11, trained SSD is more accurate with the higher IOU parameter values.
- **Sign recognition accuracy**: We evaluated our model on isoGD dataset. As Table 9 shows, our model outperforms the state-of-the-art result on this dataset.

Table 8 Results of the proposed models on isoGD dataset

Model	Iteration	Accuracy
SSD+VGG16+LSTM	10000	82.78
SSD+VGG16+LSTM	20000	82.94
SSD+ResNet50+LSTM	10000	83.15
SSD+ResNet50+LSTM	20000	83.45
SSD+VGG16+ESHR+LSTM	10000	83.75
SSD+VGG16+ESHR+LSTM	20000	83.92
SSD+ResNet50+ESHR+LSTM	10000	84.65
SSD+ResNet50+ESHR+LSTM	20000	84.95
SSD+ResNet50+ESHR+HP+LSTM	10000	86.15
SSD+ResNet50+ESHR+HP+LSTM	20000	**86.32**
SSD+ResNet50+ESHR+HP+GRU	20000	85.96

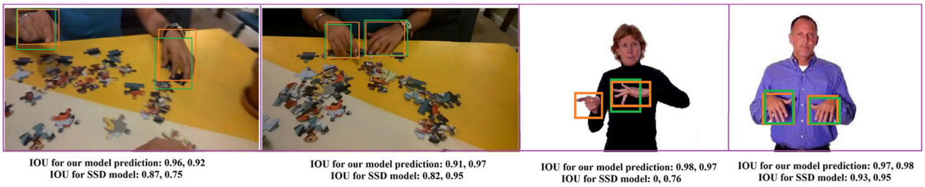

IOU for our model prediction: 0.96, 0.92
IOU for SSD model: 0.87, 0.75
IOU for our model prediction: 0.91, 0.97
IOU for SSD model: 0.82, 0.95
IOU for our model prediction: 0.98, 0.97
IOU for SSD model: 0, 0.76
IOU for our model prediction: 0.97, 0.98
IOU for SSD model: 0.93, 0.95

Fig. 11 Comparison the hand detection results of the trained SSD with original SSD. The green boxes are the detected hand boxes using original SSD. The yellow boxes are the hand detected using SSD model trained on five online sign dictionaries. IOU values show the overlap between the predicted hand box and ground-truth one

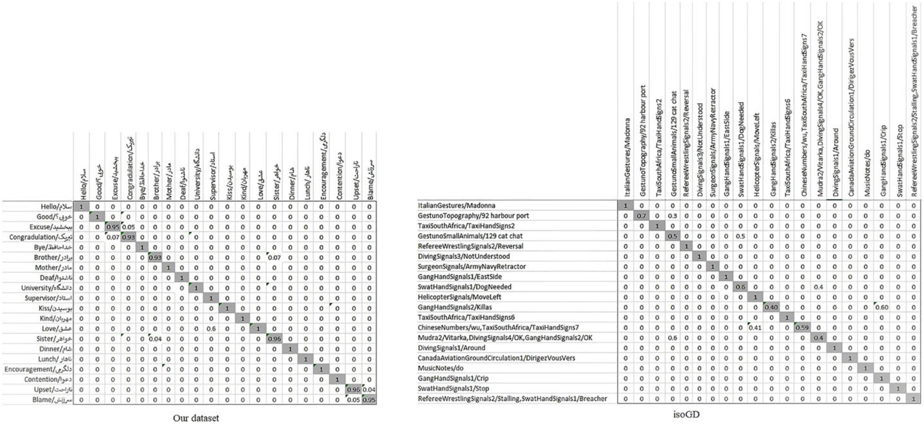
6 Discussion

Here, we analyze the results from three perspectives as follows:

- **Dataset**: While our dataset includes more samples per class in comparison with the current sign datasets, we need to extend our words domain to include more words of sign language dictionary. Also, more contributors are necessary to increase not only the generalization capability of the model but also the samples diversity for each sign.
- **Complexity and efficiency**: To have an efficient model, we tried to decrease the model complexity with preserving or improving the recognition accuracy as much as possible. We used ResNet50 model with less parameters than VGG16, as a pre-train model for spatial feature extraction. Furthermore, we tried to use GRU instead of LSTM for temporal feature extraction but the recognition accuracy of our model, including GRU, got worse. So, we skipped it in the current version of our model to make a trade-off between the model complexity and accuracy. We would try to study the possibility of using the other pre-train models with less complexity and the other models for temporal features extraction such as transformer or GRU to decrease the model complexity a little bit more. Training the SSD model on the 106’518 frames with their annotations takes about 72 hours. Also, we spent about 100 hours for LSTM training on our dataset and 207 hours on isoGD. We used the ResNet50 model only as a feature extractor and did not train it. In addition, we did not train the hand pose model because our dataset do not include the hand pose annotations. So, we used the hand pose model, as a pre-trained model for hand pose features estimation. In the test phase, our model takes about 4 hours for LSTM prediction on our dataset and 12 hours on isoGD. Prediction time of our model for each sign is approximately 2.58 second.
- **Accuracy**: Analyzing the achieved results on two datasets, our dataset and isoGD, show that our model has the higher recognition accuracy on the proposed dataset than the

Table 9 Comparison of model accuracy with state-of-the-art model on isoGD

Ref.	Model	Accuracy	Year
[5]	CNN	67.19	2016
[18]	CNN	67.71	2017
[36]	CNN	65.59	2017
[39]	CNN	60.47	2017
[20]	CNN	82.07	2018
Ours	SSD, CNN, LSTM	**86.32**	**2019**



	Hello/سلام	Good/خوبی؟	Excuse/ببخشید	Congradulation/تبریک	Bye/خداحافظ	Brother/برادر	Mother/مادر	Deaf/ناشنوا	University/دانشگاه	Supervisor/استاد	Kiss/بوسیدن	Kind/مهربان	Love/عشق	Sister/خواهر	Dinner/شام	Lunch/ ناهار	Encouragement/دلگرمی	Contention/دعوا	Upset/ناراحت	Blame/سرزنش
Hello/سلام	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Good/خوبی؟	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Excuse/ببخشید	0	0	0.95	0.05	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Congradulation/تبریک	0	0	0.07	0.93	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Bye/خداحافظ	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Brother/برادر	0	0	0	0	0	0.93	0	0	0	0	0	0	0	0.07	0	0	0	0	0	0
Mother/مادر	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
Deaf/ناشنوا	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
University/دانشگاه	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Supervisor/استاد	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
Kiss/بوسیدن	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Kind/مهربان	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Love/عشق	0	0	0	0	0	0	0	0	0	0.6	0	0	1	0	0	0	0	0	0	0
Sister/خواهر	0	0	0	0	0	0.04	0	0	0	0	0	0	0	0.96	0	0	0	0	0	0
Dinner/شام	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
Lunch/ ناهار	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
Encouragement/دلگرمی	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
Contention/دعوا	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
Upset/ناراحت	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.96	0.04
Blame/سرزنش	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.05	0.95

Our dataset

	ItalianGestures/Madonna	GestunoTopography/92 harbour port	TaxiSouthAfrica/TaxiHandSigns2	GestunoSmallAnimals/129 cat chat	RefereeWrestlingSignals2/Reversal	DivingSignals3/NotUnderstood	SurgeonSignals/ArmyNavyRetractor	GangHandSignals1/EastSide	SwatHandSignals1/DogNeeded	HelicopterSignals/MoveLeft	GangHandSignals2/Killas	TaxiSouthAfrica/TaxiHandSigns6	ChineseNumbers/wu,TaxiSouthAfrica/TaxiHandSigns7	Mudra2/Vitarka,DivingSignals4/OK,GangHandSignals2/OK	DivingSignals1/Around	CanadaAviationGroundCirculation1/DirigezVousVers	MusicNotes/do	GangHandSignals1/Crip	SwatHandSignals1/Stop	RefereeWrestlingSignals2/Stalling,SwatHandSignals1/Breacher
ItalianGestures/Madonna	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GestunoTopography/92 harbour port	0	0.7	0	0.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TaxiSouthAfrica/TaxiHandSigns2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
GestunoSmallAnimals/129 cat chat	0	0	0	0.5	0	0	0	0	0.5	0	0	0	0	0	0	0	0	0	0	0
RefereeWrestlingSignals2/Reversal	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
DivingSignals3/NotUnderstood	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
SurgeonSignals/ArmyNavyRetractor	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
GangHandSignals1/EastSide	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
SwatHandSignals1/DogNeeded	0	0	0	0	0	0	0	0	0.6	0	0	0	0	0.4	0	0	0	0	0	0
HelicopterSignals/MoveLeft	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
GangHandSignals2/Killas	0	0	0	0	0	0	0	0	0	0	0.40	0	0	0	0	0	0	0.60	0	0
TaxiSouthAfrica/TaxiHandSigns6	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
ChineseNumbers/wu,TaxiSouthAfrica/TaxiHandSigns7	0	0	0	0	0	0	0	0	0	0.41	0	0	0.59	0	0	0	0	0	0	0
Mudra2/Vitarka,DivingSignals4/OK,GangHandSignals2/OK	0	0	0	0.6	0	0	0	0	0	0	0	0	0	0.4	0	0	0	0	0	0
DivingSignals1/Around	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
CanadaAviationGroundCirculation1/DirigezVousVers	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
MusicNotes/do	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
GangHandSignals1/Crip	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
SwatHandSignals1/Stop	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
RefereeWrestlingSignals2/Stalling,SwatHandSignals1/Breacher	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

isoGD

Fig. 12 Confusion matrix of our model on the proposed dataset and isoGD

isoGD, due to have a static and near homogeneous background in this dataset and different distance assumptions. Based on the experimental results, the achieved accuracy of each sign in the proposed dataset is more than 0.93. As one can see in Fig. 12, there are some challenges for hand sign recognition in ‘Excuse’, ‘Congratulation’, ‘Sister’, ‘Brother’, ‘Upset’, ‘Blame’, ‘Fight’, ‘Competition’ signs. Analyzing the false recognition shows that there are some similarities between these signs. For example, ‘Excuse’ and ‘Congratulation’, ‘Upset’ and ‘Blame’, ‘Fight’ and ‘Competition’ signs include many similar frames and maybe we need to provide more samples in these signs or more powerful features to solve these false recognition. Figure 13 shows some samples of these false recognition. Due to using a dataset with predefined distance adjustment, our model is compatible with these assumptions. So, this model could be applied to the applications with similar distance assumptions. To be compatible with different distances, we need to train the model with more data including different distance

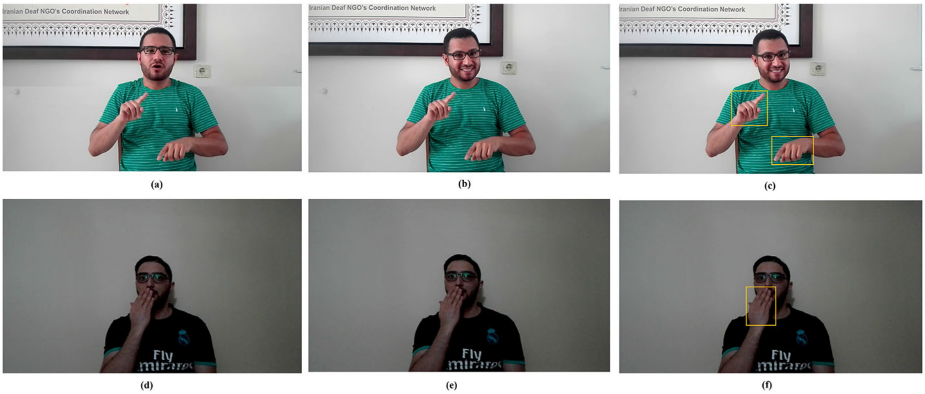


Fig. 13 Some examples of false recognition of our model. **a** Original frame of ‘Competition’ sign, **b**: Original frame of ‘Fight’ sign, **c**: Detected sign: ‘Competition’, Correct sign: ‘Fight’, **d**: Original frame of ‘Congratulation’, **e**: Original frame of ‘Excuse’ sign, **f**: Detected sign: ‘Congratulation’, Correct sign: ‘Excuse’

assumptions. Given that datasets include 100 and 249 class labels, for visualization and interpretation purposes we just select the 20 more difficult categories of each dataset for confusion matrix visualization. Figure 12 shows the confusion matrix of first 20 classes of this dataset.

To sum up, while we improved hand sign recognition accuracy and complexity, more endeavor is indispensable to provide a fast processing in an uncontrolled environment considering rapid hand motions. In this paper, we used SSD model for hand detection. While we improved the detection accuracy using large amount of annotated data, we think that it could be more improved using more data to get more robust for hand occlusion and deformation in different distance assumptions. Also, we need more data to generalize the model with more sign labels and contributors. We would like to develop this model for real-world communication of deaf and speaking disable people to facilitate their communication with themselves or the other people of society.

7 Conclusion

In this paper, we proposed a deep-based model for efficient hand sign recognition using a cascaded architecture of SSD, CNN, and LSTM from RGB videos. We trained SSD model for hand detection using annotated videos of five online sign dictionaries. We used our proposed dataset including 10’000 Persian word video samples of 10 contributors for 100 words in 10 different backgrounds. A combinational model including a CNN and LSTM with three types of spatial features has been used for hand sign recognition from detected hands of SSD model in an efficient way. We did a complete analysis to use different pretrain models, different spatial features, and different temporal-based models for sequence learning. We took ResNet50 model for feature extraction from still RGB frames. Three types of spatial features, including hand features, ESHR features, and HP features, have been fused to feed to LSTM for temporal feature extraction. The proposed model is simple yet fast and accurate that outperforms state-of-the-art results on isoGD dataset for hand sign recognition.

Acknowledgements This work is partially supported by the Spanish project TIN2016-74946-P (MINECO/FEDER,UE), CERCA Programme /Generalitat de Catalunya, ICREA under the ICREA Academia programme, and High Intelligent Solution (HIS) company of Iran. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan XP GPU used for this research. Also, we would like to thank two deaf centers of Iran (Semnan and Tehran) and the Computer Vision Center (CVC) of Spain for their collaborations.

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

Compliance with Ethical Standards

Conflict of interests The authors certify that they have no conflict of interest.

References

1. Acton B, Koum J (2009) Yahoo.www.whatsapp.com

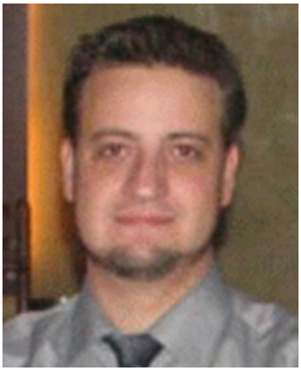
2. Chai X, Guang L, Lin Y, Xu Zh, Tang Y, Chen X, Zhou M (2013) Sign language recognition and translation with kinect. In: IEEE International conference on automatic face and gesture recognition (FG2013). April 22–26. Shanghai
3. Chen Ch, Zhang B, Zhenjie H, Jiang J, Liu M, Yang Y (2017) Action recognition from depth sequences using weighted fusion of 2D and 3D auto-correlation of gradients features. Multimedia Tools and Applications
4. Cooper H, Ong W-J, Pugeault N, Bowden R (2012) Sign language recognition using sub-units. J Mach Learn Res 13:2205–2231
5. Duan J, Zhou Sh, Wan J, Guo X, Li SZ (2016) Multi-modality fusion based on consensus-voting and 3D convolution for isolated gesture recognition, arXiv:1611.06689v2
6. El Khattabi Z, Tabii Y, Benkaddour A (2015) Video summarization: techniques and applications. Int J Comput Inform Eng 4:9
7. Forster et al (2012) WTH-PHOENIX v1 - German sign language RWTH-PHOENIX v2
8. Ge L, Liang H, Yuan J, Thalmann D (2018) Robust 3D hand pose estimation in single depth images: from single-view CNN to multi-view CNNs. IEEE Transactions on Image Processing
9. Goodwyn S, Acredolo L, Brown C (2000) Impact of symbolic gesturing on early language development. Nonverbal Behavior, 81–103. https://www.babysignlanguage.com/dictionary/?v=04c19fa1e772
10. He K, Zhang X, Ren Sh, Sun J (2016) Deep residual learning for image recognition. CVPR
11. Jameson L et al (2004) American Sign Language
12. Kang B, Tripathi S, Nguyen TQ (2015) Real-time sign language fingerspelling recognition using convolutional neural networks from depth map. In: 3rd IAPR Asian conference on pattern recognition (ACPR)
13. Kapuscinski T, Oszust M, Wysocki M, Warchol D (2015) Recognition of hand gestures observed by depth cameras. International Journal of Advanced Robotic Systems
14. Kim S, Ban Y, Lee S (2017) Tracking and classification of in-air hand gesture based on thermal guided joint filter. Sensors
15. Koller O, Forster J, Hermann N (2015) Continuous sign language recognition: towards large vocabulary statistical recognition systems handling multiple signers. Comput Vis Image Underst, 108–125
16. Le TH, Jaw DW, Lin ICh, Liu HB, Huang ShCh (2018) An efficient hand detection method based on convolutional neural network. In: The 7th IEEE international symposium on next-generation electronics
17. Liu W, Anguelov D, Erhan D, Szegedy Ch, Reed S, Fu ChY, Berg AC (2016) SSD: single shot MultiBox detector. ECCV, 21–37
18. Miao Q, Li Y, Ouyang W, Ma Z, Xu X, Shi W, Cao X, Liu Z, Chai X, Liu Z et al (2017) Multimodal gesture recognition based on the resc3d network. In: Proceedings of the IEEE conference on computer vision and pattern recognition
19. Miller J, Winn B, Winn J (2019) Signing savvy. Online dictionary
20. Narayana P, Beveridge JR, Bruce AD (2018) Gesture recognition: focus on the hands. CVPR, 5235–5244
21. Neverova N, Wolf Ch, Taylor GW, Nebout F (2014) Hand segmentation with structured convolutional learning. In: Asian conference on computer vision (ACCV) 2014: computer vision, pp 687–702
22. Ong WJ, Cooper H, Pugeault N, Bowden R (2012) Sign language recognition using sequential pattern trees. CVPR
23. Oszust M, Wysocki M (2013) Polish sign language words recognition with Kinect. In: 6th International conference on human system interactions (HSI)
24. Pagebites Inc. (2019) United States. www.imo.com
25. Pugeault N, Bowden R (2011) Spelling it out: real-time ASL fingerspelling recognition. In: Proceedings of the 1st IEEE workshop on consumer depth cameras for computer vision, jointly with ICCV'2011
26. Rastgoo R, Kiani K, Escalera S (2018) Multi-modal deep hand sign language recognition in still images using restricted Boltzmann machine. Entropy 20:809
27. Rastgoo R, Kiani K, Escalera S (2020) Hand sign language recognition using multi-view hand skeleton. Expert Syst Appl 150:113336. https://doi.org/10.1016/j.eswa.2020.113336
28. Ren S, He K, Girshick R, Sun J (2015) Faster R-CNN: towards real-time object detection with region proposal networks. NIPS
29. Ronchetti F, Quiroga F, Estrebou C, Lanzarini L (2016) Handshape recognition for Argentinian sign language using ProbSom. JCS-T
30. Ronchetti F, Quiroga F, Estrebou C, Lanzarini LC, Rosete A (2016) LSA64: an Argentinian sign language dataset. Congreso Argentino de Ciencias de la Computación (CACIC 2016)
31. Scogin J (2008) Texas math sign language dictionary. http://www.tsdvideo.org/about.php
32. Simon T, Joo H, Matthews I, Sheikh Y (2017) Hand keypoint detection in single images using multi-view bootstrapping. arXiv:1704.07809

33. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556v6
34. Sun A, Wei Y, Liang S, Tang X, Sun J (2015) Cascaded hand pose regression. CVPR, 824–832
35. Thangali A, Nash J, Sclaroff S, Neidle C (2011) Exploiting phonological constraints for handshape inference in ASL video. CVPR
36. Wang H, Wang P, Song Z, Li W (2017) Large-scale multimodal gesture recognition using heterogeneous networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition
37. William V (2013) American sign language. William Vicars Publisher, http://www.lifeprint.com/index.htm
38. Yan Sh, Xia Y, Smith JS, Lu W, Zhang B (2017) Multi-scale convolutional neural networks for hand detection. Applied Computational Intelligence and Soft Computing
39. Zhang L, Zhu G, Shen P, Song J, Shah SA, Bennamoun M (2017) Learning spatiotemporal features using 3dcnn and convolutional lstm for gesture recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition
40. Zhou Y, Lu J, Lin X, Sun Y, Ma X (2018) HBE: hand branch ensemble network for real-time 3D Hand Pose Estimation. ECCV
41. Zimmermann Ch, Brox T (2017) Learning to estimate 3D hand pose from single RGB images. ICCV

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Razieh Rastgoo received the B.Sc. Degree in Computer Engineering, Hardware, from Shiraz University of Iran. Also, she achieved her M.Sc. in Artificial Intelligence. She currently is a Ph.D. student at Semnan University in Iran. Furthermore, she is a researcher at HIS Company in Iran and HUPBA Lab in University of Barcelona, Spain. Her interest areas are: Artificial Intelligence, Machine Learning, Deep Learning, Computer Vision, Sign Language, Pattern Recognition, Natural Language Processing (NLP), Smart Grids, Routing Protocols, Computer Networks.

Kourosh Kiani received the B.Sc. and M.Sc., degrees in Electrical Engineering from Delft University of Technology at Netherlands in 1993 and the Ph.D. degree in Medical information from Erasmus University in Rotterdam, the Netherlands in 1997. He is currently an Assistant Professor with the Faculty of Electrical and Computer Engineering, Semnan University, Semnan, Iran. His research interests include Artificial Intelligence, Machine learning, Deep learning, Big Data, Pattern Recognition, Natural Language Processing (NLP).



Sergio Escalera obtained the P.h.D. degree on Multi-class visual categorization systems at Computer Vision Center, UAB. He obtained the 2008 best Thesis award on Computer Science at Universitat Autònoma de Barcelona. He leads the Human Pose Recovery and Behavior Analysis Group at UB, CVC, and the Barcelona Graduate School of Mathematics. He is an associate professor at the Department of Mathematics and Informatics, Universitat de Barcelona. He is an adjunct professor at Universitat Oberta de Catalunya, Aalborg University, and Dalhousie University. He has been visiting professor at TU Delft and Aalborg Universities. He is a member of the Visual and Computational Learning consolidated research group of Catalonia. He is also a member of the Computer Vision Center at UAB. He is series editor of The Springer Series on Challenges in Machine Learning. He is Editor-in-Chief of American Journal of Intelligent Systems and editorial board member of more than 5 international journals. He is vice-president of ChaLearn Challenges in Machine Learning, leading ChaLearn Looking at People events. He is co-creator of Codalab open source platform for challenges organization. He is co-founder of PhysicalTech and Care Respite companies. He is also member of the AERFAI Spanish Association on Pattern Recognition, ACIA Catalan Association of Artificial Intelligence, INNS, and Chair of IAPR TC-12: Multimedia and visual information systems. He has different patents and registered models. He has published more than 250 research papers and participated in the organization of scientific events, including CCIA04, ICCV11, CCIA14, AMDO16, FG17, NIPS17, NIPS18, FG19, and workshops at ICCV, ICMI, ECCV, CVPR, ICPR, NIPS. He has been guest editor at JMLR, TPAMI, IJCV, TAC, PR, MVA, JIVP, Expert Systems, and Neural Comp. and App. He has been area chair at WACV16, NIPS16, AVSS17, FG17, ICCV17, WACV18, FG18, BMVC18, NIPS18, FG19 and competition and demo chair at FG17, NIPS17, NIPS18, ECMLPKDD19 and FG19. His research interests include, statistical pattern recognition, affective computing, and human pose recovery and behavior understanding, including multimodal data analysis, with special interest in characterizing people: personality and psychological profile computing.