ChaLearn Looking at People: IsoGD and ConGD Large-Scale RGB-D Gesture Recognition

Jun Wan, *Senior Member, IEEE*, Chi Lin, *Member, IEEE*, Longyin Wen, Yunan Li, Qiguang Miao, *Senior Member, IEEE*, Sergio Escalera, *Member, IEEE*, Gholamreza Anbarjafari, *Senior Member, IEEE*, Isabelle Guyon, Guodong Guo, *Senior Member, IEEE*, and Stan Z. Li, *Fellow, IEEE*

***Abstract*—The ChaLearn large-scale gesture recognition challenge has run twice in two workshops in conjunction with the International Conference on Pattern Recognition (ICPR) 2016 and International Conference on Computer Vision (ICCV) 2017, attracting more than 200 teams around the world. This challenge has two tracks, focusing on isolated and continuous gesture recognition, respectively. It describes the creation of both benchmark datasets and analyzes the advances in large-scale gesture recognition based on these two datasets. In this article, we discuss the challenges of collecting large-scale ground-truth annotations of gesture recognition and provide a detailed analysis of the current methods for large-scale isolated and continuous gesture recognition. In addition to the recognition rate and mean Jaccard index (MJI) as evaluation metrics used in previous challenges, we introduce the corrected segmentation rate (CSR) metric to evaluate the performance of temporal segmentation for continuous gesture recognition. Furthermore, we propose a bidirectional long short-term memory (Bi-LSTM) method, determining video division points based on skeleton points. Experiments show that the proposed Bi-LSTM outperforms state-of-the-art methods with an absolute improvement of 8.1% (from 0.8917 to 0.9639) of CSR.**

***Index Terms*—Bidirectional long short-term memory (Bi-LSTM), gesture recognition, RGB-D.**

Manuscript received December 16, 2019; revised April 29, 2020; accepted July 19, 2020. Date of publication August 20, 2020; date of current version May 19, 2022. This work was supported in part by the Chinese National Natural Science Foundation Projects under Grant 61961160704 and Grant 61876179; in part by the Key Project of the General Logistics Department under Grant ASW17C001; in part by the Science and Technology Development Fund of Macau under Grant 0010/2019/AFJ and Grant 0025/2019/AKP; in part by the Spanish Project PID2019-105093GB-I00 (MINECO/FEDER, UE) and CERCA Programme/Generalitat de Catalunya); in part by ICREA through the ICREA Academia Programme; and in part by the Estonian Centre of Excellence in IT (EXCITE) funded by the European Regional Development Fund. This article was recommended by Associate Editor D. Goldgof. *(Corresponding author: Jun Wan.)*

Jun Wan is with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, and University of Chinese Academy of Sciences, Beijing 100190, China (e-mail: jun.wan@ia.ac.cn).

Chi Lin is with the Department of Computer Science, University of Southern California, Los Angeles, CA 90089 USA (e-mail: linchi@usc.edu).

Longyin Wen is with JD Finance, Mountain View, CA 94043 USA (e-mail: longyin.wen@jd.com).

Yunan Li and Qiguang Miao are with the School of Computer Science and Technology, Xidian University, Xi'an 710071, China, and also with the Xi'an Key Laboratory of Big Data and Intelligent Vision, Xidian University, Xi'an 710071, China (e-mail: yunanli@xidian.edu.cn; qgmiao@xidian.edu.cn).

Sergio Escalera is with the Computer Vision Center, Universitat de Barcelona, 08007 Barcelona, Spain (e-mail: sergio@maia.ub.es).

Gholamreza Anbarjafari is with iCV Lab, Institute of Technology, University of Tartu, 50090 Tartu, Estonia, also with PwC Finland, 00180 Helsinki, Finland, and also with the Faculty of Engineering, Hasan Kalyoncu University, 27100 Gaziantep, Turkey (e-mail: shb@ut.ee).

Isabelle Guyon is with ChaLearn, San Francisco, CA 94115 USA, and also with University Paris-Saclay, 91190 Saint-Aubin, France (e-mail: guyon@chalearn.org).

Guodong Guo is with the Institute of Deep Learning, Baidu Research, Beijing 100193, China, and National Engineering Laboratory for Deep Learning Technology and Application, Beijing 100193, China (e-mail: guoguodong01@baidu.com).

Stan Z. Li is with the Westlake University, Hangzhou 310024, China, and also with the Macau University of Science and Technology, Macau, China (e-mail: stan.zq.li@westlake.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2020.3012092.

Digital Object Identifier 10.1109/TCYB.2020.3012092

I. Introduction

HUMAN action and gesture recognition have received a lot of attention from the computer vision community. In the past few years, several famous action and gesture datasets have been released, such as the NTU RGB+D dataset [7], MSR-Action3D dataset [8], CAD-60 [9] and CAD-120 dataset [10], DHG2016 dataset [11], SHREC17 dataset [12], RGBD-HuDaAct dataset [13], SYSU 3-D human–object interaction (3DHOI) dataset [14], HMDB51 dataset [15], UCF101 dataset [16] and Kinetics datasets [17], CGD [1], multimodal gesture dataset [2], Sheffield gesture dataset [4], NIVIDIA gesture dataset [5], and EgoGesture dataset [6]. These datasets pushed the advance of the state-of-the-art research for action and gesture recognition [18], [19]. Interestingly, the data size of action datasets is much larger than gesture datasets in terms of both the amount of data and the number of classes. For example, the Kinetics dataset includes 600 action classes and 500 000 video clips while only limited data are provided for gesture recognition (i.e., the Sheffield dataset contains about ten classes, 1080 videos). The main reasons for this are: 1) actions can be more easily captured than gestures, such as the Kinects dataset from YouTube videos; 2) the gesture can be seen as a semiotic sign highly dependent on the cultural context (i.e., Chinese number) while the action is goal-directed motion sequence (i.e., play football). Therefore, if a large amount of gestures is needed, it requires a lot of human-labor costs; and 3) actions tend to focus on the body information with large motions (such as hugging and sports) while gestures are produced as part of deliberate actions and signs, involving the motion of the

up body, especially the arms, hands, and fingers. Furthermore, facial expressions are also considered to be involved in gesture recognition. The previous leads to high complexity of collecting large gesture recognition datasets. In addition, because of the different nature between signs and gestures, there is no guarantee existing action datasets are suitable for training systems for gesture-based recognition scenarios, such as sign language recognition and human–computer interaction, where not only the body part motion is needed but also the semantic language and facial expressions. Due to the above reasons, one can see that the released gesture datasets in Table I are very limited in size. This hinders further developments of deep learning-based methods in gesture recognition.

To this end, we present two large-scale datasets with RGB-D video sequences, namely, the ChaLearn-isolated gesture dataset (IsoGD) and the continuous gesture dataset (ConGD) for the tasks of isolated and continuous gesture recognition, respectively. Both of them consist of more than 47 000 gestures fallen into 249 classes performed by 21 performers. Besides, we organized two ChaLearn large-scale gesture recognition challenge workshops in conjunction with the International Conference on Pattern Recognition (ICPR) 2016 [20] and International Conference on Computer Vision (ICCV) 2017 [21]. The datasets allowed for the development and comparison of different algorithms, and the competition and workshop provided a way to track the progress and discuss advantages and disadvantages learned from the most successful and innovative entries.

The main contributions are summarized as follows.

1) We discuss the challenges of creating two large-scale gesture benchmark datasets, namely, the IsoGD and ConGD, and highlight developments in both isolated and continuous gesture recognition fields by creating the benchmark and holding the challenges. We analyze the submitted results in both challenges and review the published algorithms in the last three years.
2) A new temporal segmentation algorithm called the bidirectional long short-term memory (Bi-LSTM) segmentation network is proposed, which is used to determine the start and end frames of each gesture in the continuous gesture video. Compared with the existing methods, the main advantage of the proposed method is to avoid the need for prior assumptions.
3) A new evaluation metric called the corrected segmentation rate (CSR) is introduced and used to evaluate the performance of temporal segmentation. Compared with the published methods, the proposed Bi-LSTM method improves state-of-the-art results. The superiority of temporal segmentation is about 8.1% (from 0.8917 to 0.9639) by CSR on the testing sets of the ConGD dataset.

The remainder of this article is organized as follows. We describe datasets, evaluation metrics, and organized challenges in Section II. In Section III, we review the state-of-the-art methods focusing on both datasets. We propose a new algorithm for temporal segmentation in Section IV and present experimental results on the two proposed datasets in Section V. Finally, we conclude this article in Section VI.

TABLE I
COMPARISONS WITH RGB-D GESTURE DATASETS

Dataset	Total gestures	Gesture labels	Avg. samp. per cls.	Train samp. (per cls.)
CGD [1], 2011	54,000	>200	10	8∼12 (1-1-1)
Multi-modal Gesture Dataset [2], 2013	13,858	20	692	7,754
ChAirGest [3], 2013	1,200	10	120	-
Sheffield Gesture Dataset [4], 2013	1,080	10	108	- -
NIVIDIA Dataset [5], 2016	1,532	25	-	1,050
EgoGesture Dataset [6], 2017	24,161	83	291	14416
IsoGD (Ours)	47,933	249	192	35,878
ConGD (Ours)	47,933	249	192	30,442

II. DATASET INTRODUCTION AND CHALLENGE TASKS

A. Motivation

Benchmark datasets can greatly promote the research developments in their respective fields. For example, the ImageNet Large-Scale Visual-Recognition Challenge [22] (ILSVRC) is held every year from 2010 to 2017, which includes several challenging tasks, including image classification, single-object localization, and object detection. The dataset presented in this challenge contains 1000 object classes with approximately 1.2 million training images, 50 000 validation images, and 100 000 testing images, which greatly promotes the development of new techniques, particularly those based on deep learning architectures, for image classification and object localization. Several other datasets have been also designed to evaluate different computer vision tasks, such as human pose recovery, action and gesture recognition, and face analysis [2], [23], [24].

Nevertheless, there are very few annotated datasets with a large number of samples and gesture categories for the task of RGB-D gesture recognition. Table I lists the publicly available RGB-D gesture datasets released from 2011 to 2017. Most datasets include less than 20 gesture classes (e.g., [3] and [4]). Although the CGD dataset [1] has about 54 000 gestures, it is designed for the one-shot learning task (only one training sample per class). The multimodal gesture dataset [2], [23] contains about 13 000 gestures with 387 training samples per class, but it only has 20 classes.

In order to provide the community with a large dataset for RGB-D gesture recognition, here we take benefit of the previous CGD dataset [1] by integrating all gesture categories and samples to design two new large RGB-D datasets for gesture spotting and classification. In Table I, the new IsoGD and ConGD datasets show a significant increase in size in terms of both the number of categories and the number of samples in comparison to state-of-the-art alternatives.

B. Dataset Introduction

As previously mentioned, our datasets were derived from the CGD dataset [1] which was designed for the "one-shot learning" task. The CGD dataset contained 540 batches (or subfolders), and all batches of CGD had 289 gestures from 30 lexicons and a large number of gestures in total (54 000 gestures in about 23 000 RGB-D video clips), which makes it a very valuable material to carve out different tasks. This is what

TABLE II
SUMMARY OF ISOGD AND CONGD DATASETS

Sets	the IsoGD dataset			the ConGD dataset		
	#gestures	#videos	#subjects	#gestures	#videos	#subjects
Train	35878	35878	17	30442	14134	17
Valid.	5784	5784	2	8889	4179	2
Test	6271	6271	2	8602	4042	2
All	47933	47933	21	47933	22535	21

we did by creating two large RGB-D gesture datasets: 1) the IsoGD[1] and 2) ConGD[2] datasets. For detailed information about gesture annotation, refer to [25]. Finally, we obtained 249 unique gesture labels from 30 gesture lexicons and 47 933 gestures in 22 535 RGB-D videos.

The detailed lexicons are shown in Fig. 1. One can see that it has wide gesture lexicons, such as ChineseNumbers, GestunoDisaster, GangHandSignals, ItalianGestures, TrafficPoliceSignals, and so on. Each gesture lexicon has 8–13 classes. One of the intentions for this is to make it feasible for people to use a branch of these lexicons to train their own model for their specific application. In Fig. 1, the pink boxes indicate different gesture classes in the IsoGD and ConGD datasets. The gray boxes are not used in our datasets because the number of samples from the MotorcycleSignals is very limited. In order to keep data balance, we omitted this lexicon. Other boxes with the same color indicate the same gesture we have merged from the CGD dataset. The details of the merged gesture classes ID in both IsoGD and ConGD datasets are released.[3]

C. Dataset Statistics

The statistical information is shown in Table II. For the ConGD dataset, it includes 47 933 RGB-D gestures in 22 535 RGB-D videos. Each RGB-D video can represent one or more gestures, and there are 249 gestures labels performed by 21 different individuals. For the IsoGD dataset, we split all videos of the ConGD dataset into isolated gestures, obtaining 47 933 gestures. Each RGB-D video represents one gesture instance, having 249 gestures labels performed by 21 individuals.

D. Evaluation Metrics

For both datasets, we provide training, validation, and test sets. In order to make it more challenging, all three sets include data from different subjects, which means the gestures of one subject in validation and test sets will not appear in the training set. According to [25], we introduced the recognition rate r and mean Jaccard index (MJI) $\bar{J}_{\text{acc}}$ as the evaluation criteria for the IsoGD and ConGD datasets, respectively.

For continuous gesture recognition, the MJI $\bar{J}_{\text{acc}}$ metric is commonly used as the comprehensive evaluation [2], [25]. However, it does not provide a specific assessment of either the classifier or the temporal segmentation strategy. Therefore, this metric makes it difficult to evaluate if a high-performance score is attributed to the classifier or the considered temporal segmentation strategy. For the classifier, the recognition rate can be used as the evaluation metric, which is similar to that of isolated gesture recognition. The CSR E_{CSR} is the first evaluation metric designed to evaluate the performance of temporal segmentation. The CSR E_{CSR} is based on intersection-over-union (IoU) and is defined as

$$E_{\text{CSR}}(p, l, r) = \frac{\sum_{i=0}^{n} \sum_{j=0}^{m} M(p_i, l_j, r)}{\max(n, m)} \tag{1}$$

where p is the target model's predicted segmentation for each video, which is constituted by positions of the starting and ending frames. l is the ground truth, which has the same form as p. n and m are the number of segmentation in the model's prediction and the ground truth, respectively. M is the function to evaluate whether the two sections match or not with a predefined threshold r, as described as follows:

$$M(a, b, r) = \begin{cases} 1, & \text{IoU}(a, b) \geq r \\ 0, & \text{IoU}(a, b) < r \end{cases} \tag{2}$$

where a is the segmentation result that needs evaluation and b is the ground truth. The IoU function is defined below, which is similar to its definition for object detection [26]

$$\text{IoU}(a, b) = \frac{a \cap b}{a \cup b} = \frac{\max(0, \min(a_e, b_e) - \max(a_s, b_s))}{\max(a_e, b_e) - \min(a_s, b_s)} \tag{3}$$

where a_s and a_e represent the starting frame and the ending frame of the segmentation a. b_s and b_e are in a manner analogous to a_s and a_e. If IoU(a, b) is greater than the threshold r, we consider that they are matched successfully.

TABLE III
SUMMARY OF PARTICIPATION FOR BOTH CHALLENGES

Challenge	#Round	#Teams[1]	#Teams[2]
Isolated Gesture Challenge	1	67	20
	2	44	17
Continuous Gesture Challenge	1	66	6
	2	39	11

[1] total number of teams; [2] teams that submitted the predicted results on test sets.

TABLE IV
SUMMARY OF THE RESULTS IN OUR CHALLENGES
(r: RECOGNITION RATE)

	Isolated Gesture Recognition				Continuous Gesture Recognition		
	rank by test	Team	evaluation		Team	evaluation	
			valid (r)	test (r)		valid (MJI)	test (MJI)
round 2	1	ASU[28]	64.40%	67.71%	ICT_NHCI[35]	0.5163	0.6103
	2	SYSU_ISEE	59.70%	67.02%	AMRL[36]	0.5957	0.5950
	3	Lostoy	62.02%	65.97%	PaFiFA[37]	0.3646	0.3744
	4	AMRL[29]	60.81%	65.59%	Deepgesture[38]	0.3190	0.316
	5	XDETVP[30]	58.00%	60.47%	-	-	-
	-	baseline [31]	49.17%	67.26%	-	-	-
round 1	1	FLiXT [32]	49.20%	56.90%	ICT_NHCI[39]	0.2655	0.2869
	2	ARML [33]	39.23%	55.57%	TARDIS [40]	0.2809	0.2692
	3	XDETVP-TRIMPS [34]	45.02%	50.93%	AMRL [41]	0.2403	0.2655
	-	baseline [25]	18.65%	24.19%	baseline [25]	0.0918	0.1464

E. Challenge Tasks

Both large-scale isolated and continuous gesture challenges belong to the series of ChaLearn LAP events,[4] which were launched in two rounds in conjunction with the ICPR (Cancun, Mexican, December 2016) and ICCV (Venice, Italy, October 2017). This competition consisted of a development phase (June 30, 2016 to August 7, 2016 for the first round, and April 20, 2017 to June 22, 2017 for the second round) and a final

[1] http://www.cbsr.ia.ac.cn/users/jwan/database/isogd.html
[2] http://www.cbsr.ia.ac.cn/users/jwan/database/congd.html
[3] http://www.cbsr.ia.ac.cn/users/jwan/database/GestureLexiconsID.pdf
[4] http://chalearnlap.cvc.uab.es/

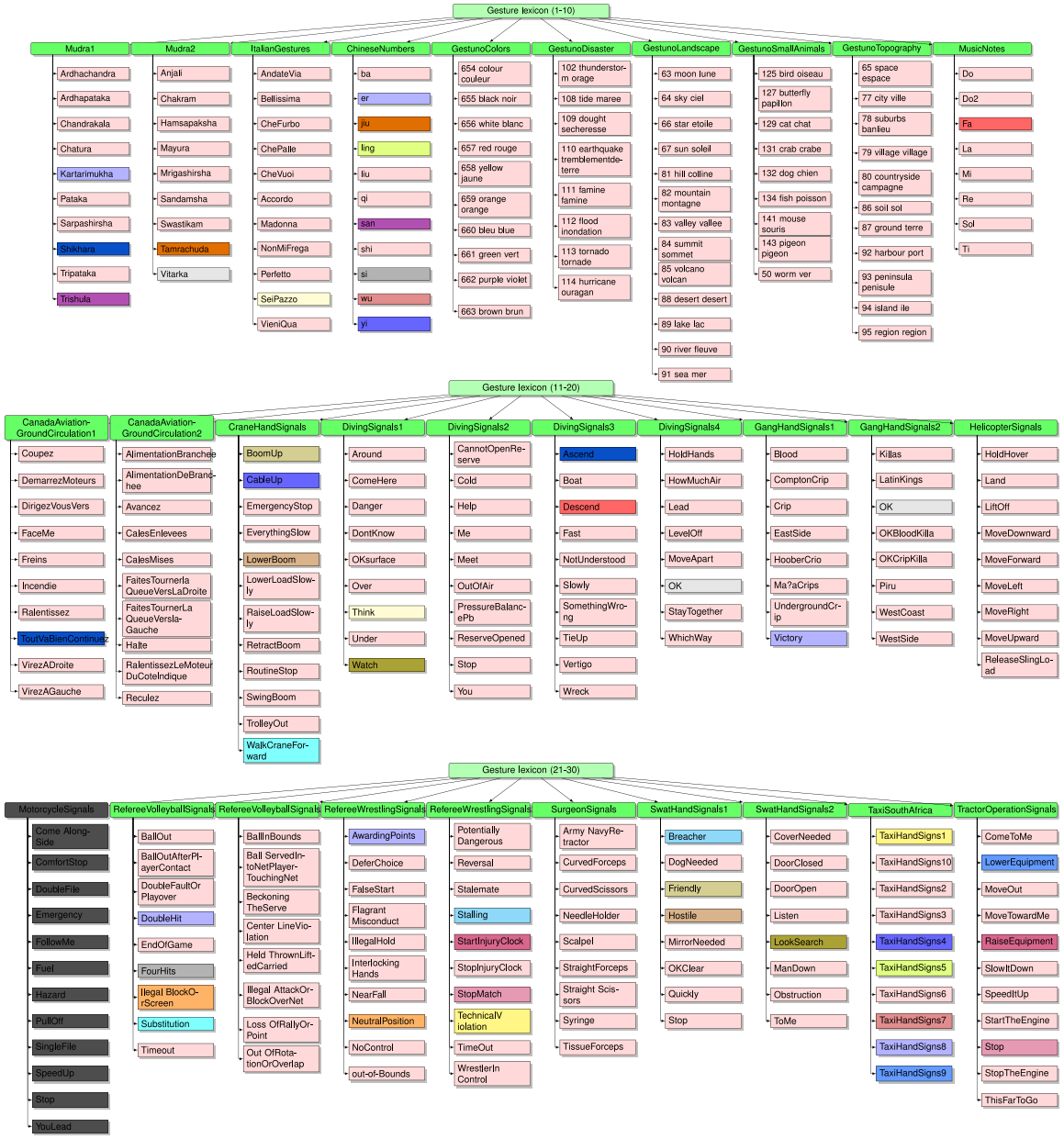


Fig. 1. Gesture lexicons in IsoGD and ConGD. The pink boxes indicate different gesture classes. The gray boxes are not used in our datasets because the number of samples from the MotorcycleSignals is very limited. In order to keep data balance, we omitted this lexicon. Other boxes with the same color indicate the same gesture we have merged from CGD.

evaluation phase (August 7, 2016 to August 17, 2016 for the first round, and June 23, 2017 to July 2, 2017 for the second round). Table III shows the summary of the participation for both gesture challenges. The total number of registered participants of both challenges is more than 200, and 54 teams have submitted their predicted results.

For each round, training, validation, and test datasets were provided. Training data were released with labels, validation data were used to provide feedback to participants in the leaderboard, and test data were used to determine the winners. Note that each track had its own evaluation metrics. The four tracks were run in the CodaLab platform.[5] The top three ranked participants for each track were eligible for prizes. The performances of winners are shown in Table IV.

[5]https://competitions.codalab.org/

III. Review of State-of-the-art Methods

In recent years, the commercialization of affordable RGB-D sensors, such as Kinect, made it available depth maps, in addition to classical RGB, which are robust against illumination variations and contain abundant 3-D structure information. Based on this technology, we created the IsoGD and ConGD datasets [25], which has been already used by several researchers to evaluate the performance of gesture recognition models. In this section, we provide a review of state-of-the-art methods using both datasets, pointing out

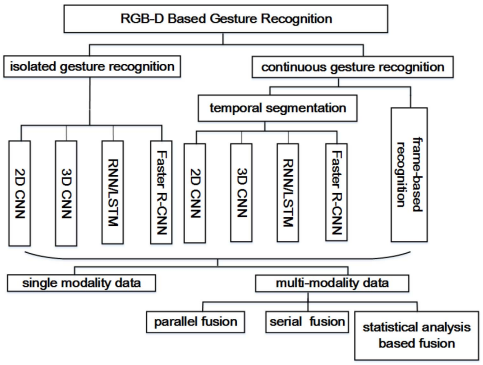


Fig. 2. Graphical representation of categories of RGB-D-based gesture recognition methods.

advantages and drawbacks, comparing them, and providing discussion for future research directions.

We summarize the recognition task into two main categories, namely, isolated and continuous gesture recognition, which are shown in Fig. 2. For the task of isolated gesture recognition, 2-D CNN-based methods [41], [42] learn spatial features while 3-D CNN-based methods [27]–[29], [31], [33], [34], [39], [43] learn spatiotemporal features. In order to obtain a tradeoff between highly discriminative features (such as 3-D spatiotemporal features) and computation complexity, 2-D CNN over dynamic images encoding spatiotemporal information have been explored in [21], [28], [32], [35], [40], and [44]. Recurrent neural networks (RNNs) [45] or its variant long short-term memory (LSTM) [46] have been also applied to analyze sequential information in videos [29], [35], [37], [38], [43]. However, the task of continuous gesture recognition involves additional challenges to the ones of classical isolated gesture recognition. In continuous gesture recognition, there may be several gestures in a video to be recognized in time. This task uses to be addressed either in a frame-by-frame fashion [36], [39] or by a temporal segmentation strategy [34], [35], [38], [40]. A detailed comparison of methods and their basic features evaluated on isolated and continuous gesture recognition datasets is shown in Tables V and VI, respectively. Below, we briefly review common techniques used in gesture recognition, such as preprocessing strategies, CNN models, and fusion strategies [47].

A. Preprocessing

In the case of the RGB modality, it is highly affected by illumination changes, while the depth modality is insensitive to illumination variations, though it may present some noisy readings based on environmental factors, such as surface reflections. In order to overcome previous issues, Miao *et al.* [27] implemented the Retinex theory [60] to normalize illumination of RGB videos, and used a median filter to denoise depth maps. Asadi-Aghbolaghi *et al.* [61] utilized a hybrid median filter and an inpainting technique to enhance depth videos.

The second category of preprocessing is based on frame unification and video calibration. The reason for frame unification is to fix the same dimension for all inputs in CNNs. After a statistical analysis of frame number distribution of training data on the IsoGD dataset, Li *et al.* [31] fixed the frame number of each clip as 32 to minimize the loss of motion

TABLE V
STATE-OF-THE-ART METHODS REVIEW ON ISOGD

Method	pre-processing	model	fusion strategy	modality of data	evaluation	
					valid	test
Wan *et al.* [50], [25]'16	/	MFSK+BoVW	SVM	RGB-D	18.65%	24.19%
Li *et al.* [32] '16	32-frame sampling	C3D	SVM	RGB-D	49.20%	56.90%
Wang *et al.* [33] '16	bidirectional rank pooling	VGG-16	score fusion	depth (DDI+ DDNI+DDMNI)	39.23%	55.57%
Zhu *et al.* [34] '16	32-frame sampling	pyramidal C3D	score fusion	RGB-D	45.02%	50.93%
Zhu *et al.* [44] '17	32-frame sampling	C3D, convLSTM	score fusion	RGB-D	51.02%	/
Wang *et al.* [51] '17	calibration	AlexNet	score fusion	RGB-D(SFAM)	36.27%	/
Li *et al.* [45] '17	32-frame sampling	C3D	SVM	RGB-D flow	54.50%	60.93%
Miao *et al.* [28] '17	Retinex, median filter, 32-frame sampling	ResC3D	SVM	RGB-D flow	64.40%	67.71%
Wang *et al.* [36] '17	bidirectional rank pooling	convLSTM, Resnet-50, C3D	score fusion	RGB-D saliency	60.81%	65.59%
Zhang *et al.* [30] '17	32-frame sampling	convLSTM, C3D	score fusion	RGB-D flow	58.00%	60.47%
Zhang *et al.* [46] '17	/	AlexNet	score fusion	depth (eDMM+SPM)	36.63%	43.91%
Duan *et al.* [31] '17	32-frame sampling	2S CNN, C3D	score fusion	RGB-D saliency	49.17%	67.26%
Hu *et al.* [52] '18	32-frame sampling	DNN	adaptive hidden layer	RGB-D	54.14%	/
Wang *et al.* [53] '18	bidirectional rank pooling	c-ConvNet	score fusion	RGB-D (VDI+DDI)	44.80%	/
Lin *et al.* [54] '18	32-frame sampling	Skeleton LSTM, C3D	adaptive weight fusion	RGB-D Skeleton	64.34%	68.42%
Wang *et al.* [55] '18	bidirectional rank pooling	convLSTM, Resnet-50, C3D	score fusion	RGB-D	43.72%	59.21%
Narayana *et al.* [56] '18	10-frame sliding window	ResNet-50	Stacked attention models	RGB-D Flow	80.96%	82.07%
Zhu *et al.* [57] '19	hand segment.	shape representaiton	DTW	RGB-D	-	60.12%
Zhu *et al.* [58] '19	-	Res3D, ConvLSTM, MobileNet	-	RGB	61.05%	-
Li *et al.* [59] '19	32-frame sampling	attention-based ResC3D	pyramidal strategy	RGB-D	-	68.14%

TABLE VI
STATE-OF-THE-ART METHODS REVIEW ON CONGD. OUR METHOD ACHIEVES THE BEST PERFORMANCE UNDER METRICS OF MJI AND CSR@IOU = 0.7 (THE HIGHER THE BETTER)

Method	preproces.	model	fusion strategy	modality of data	evaluation			
					valid		test	
					MJI	CSR	MJI	CSR
Wan *et al.* [50], [25]'16	temporal segment.	MFSK+ BoVW	SVM	RGB-D	0.0918	/	0.1464	/
Wang *et al.* [41] '16	temporal segment.	CNN	/	depth (IDMM)	0.2403	0.6636	0.2655	0.7520
Chai *et al.* [39] '16	temporal segment.	2S-RNN, Fast. R-CNN, LSTM	/	RGB-D	0.2655	/	0.2869	0.3213
Camgoz *et al.* [40] '16	/	C3D	/	RGB	0.3430	/	0.3148	0.6603
Pigou *et al.* [38] '17	/	Resnet, LSTM	score fusion	gray-scale	0.3190	0.6159	0.3164	0.6241
Camgoz *et al.* [37] '17	temporal segment.	C3D	/	RGB	0.3806	0.8213	0.3744	0.8254
Wang *et al.* [29] '17	bi-rank pooling, temporal segment.	convLSTM, C3D	score fusion	RGB-D (DDI+DRI)	0.5957	0.6636	0.5950	0.7520
Liu *et al.* [35] '17	calib, 32-frame sampling, temporal segment.	C3D, Faster R-CNN	SVM	RGB-D	0.5163	0.9034	0.6103	0.8917
Zhu *et al.* [60] '18	temporal segment.	3D CNN, ConvLSTM	score fusion	RGB-D Flow	0.5368	0.8553	0.7163	0.8776
Hoang *et al.* [61] '19	temporal segment.	3D CNN, Bi-LSTM [54]	score fusion	RGB-D, skeleton	-	-	0.5523	-
Bi-LSTM (Ours)	temporal segment.	C3D, Bi-LSTM [54]	adapt. score fusion [54]	skeleton via CPM	**0.6830**	**0.9668**	**0.7179**	**0.9639**

path in the temporal dimension. The same criterion has been used by most subsequent methods [27], [29], [31], [34], [43]. Meanwhile, although the RGB and depth videos are captured concurrently by many devices, for example, the Kinect sensor, RGB, and depth images are not accurately registered spatially. Such a spatial misalignment may affect multimodality fusion. Therefore, Wang *et al.* [49] proposed a self-calibration strategy based on a pinhole model to register multimodal data. Similarly, Asadi-Aghbolaghi *et al.* [61] exploited the intrinsic and extrinsic parameters of cameras to warp the RGB image to fit the depth one.

For continuous gesture recognition, temporal segmentation is widely applied to split continuous gestures into several isolated gestures. For example, Chai *et al.* [38] first took such a segmentation strategy for continuous gesture recognition. It assumes all gestures begin and end with performers' hands down. Then, the video can be characterized as successive gesture parts and transition parts. A similar idea is used in [34], [35], and [40]. Camgoz *et al.* [36] conducted such a temporal segmentation in a different way. They treat the segmentation process as a feature to learn and use the likelihood to split the videos into multiple isolated segments, which is done by localizing the silence regions, where there is no motion. Then, the motion of the hand palm and finger movements are fed into a Bi-LSTM network for gesture spotting [59]. Although the work of Hoang *et al.* [59] uses Bi-LSTM, similar to us, our architecture and inputs are different, and we achieve higher recognition performance, as shown in Table VI.

B. Deep Learning-Based Methods

Due to the overwhelming advantage of deep learning versus handcrafted feature-based methods, next, we briefly review deep learning-based methods, which are grouped into 2-D CNNs, 3-D CNNs, and RNN/LSTM for gesture recognition, and Faster R-CNN for hand detection.

2-D CNNs: 2-D CNNs, such as AlexNet [41], VGG [42], and ResNet [62], have shown great performance dealing with still-image recognition tasks. There are several methods [32], [49], [63] that implement the 2-D CNN to extract spatial features. In order to extend 2-D CNN to consider temporal information, Wang *et al.* [32], [35] used rank pooling [64] to generate dynamic depth images (DDIs) and computed dynamic depth normal images (DDNIs) and dynamic depth motion normal images (DDMNIs) to wrap both the motion information and the static posture in an image. The counterpart work of Wang *et al.* [51] uses RGB videos to generate visual dynamic images (VDIs). The work of Wang *et al.* [28] extends the DDIs for both body and hand level representation, which are called body-level DDIs (BDDIs) and hand-level DDIs (HDDIs), respectively. Zhang *et al.* [44] used an enhanced depth motion map (eDMM) to describe depth videos and a static pose map (SPM) for postures. Then, two CNNs are used to extract features from these representations. Wang *et al.* [49] used the scene flow vector, which is obtained by registered RGB-D data, as a descriptor to generate an action map, which is subsequently fed into AlexNet for classification.

3-D CNNs: 3-D CNNs such as C3D [65] were proposed to extend 2-D CNNs to compute spatiotemporal features. Li *et al.* [31], [78] utilized 3-D CNN to extract features from RGB-D, saliency, and optical-flow videos. Zhu *et al.* [33], [43] proposed a pyramid 3-D CNN model, in which the videos are divided into three 16 frame clips, performing prediction in each of them. Final recognition is obtained by means of score fusion. Such a pyramid 3-D CNN model is also employed by [28], [29], and [35]. Liu *et al.* [34] and Wang *et al.* [35] extended 3-D CNN in a similar fashion for continuous gesture recognition, first splitting the continuous gestures into isolated ones using temporal segmentation. In [36], 3-D CNN is used in a framewise fashion, with the final classification given by posterior estimation after several iterations.

RNN/LSTM: The RNN [45], or its variation, LSTM [46] is a kind of network where connections between units form a directed cycle. The special structure of RNN-like models allows for sequential input analysis. Chai *et al.* [38] used two streams of RNN to represent features of RGB-D videos and used LSTM to model the context. Pigou *et al.* [37] first used a ResNet to extract features of gray-scale video, and then used a bidirectional LSTM [66] to process both temporal directions. Zhu *et al.* [43] used convolutional LSTM (ConvLSTM) with 3-D CNN input to model the sequential relations between small video clips. The 3-D CNN + LSTM scheme is also employed in [29], [35], and [52].

Faster R-CNN: The faster R-CNN [67] was initially proposed for object detection tasks. Some works in gesture recognition used object detection algorithms for hand detection [21], [28], [34], [35], [38]. Chai *et al.* [38] used faster R-CNN to detect hands for recognizing begin–end gesture instances. The same strategy is applied in related works [34], [35]. Some methods further combined global and local hand regions to boost recognition performance [21], [28].

Attention Models: Some attention-aware methods [54], [68] have been applied for gesture recognition. For example, Narayana *et al.* [54] proposed a focus of the attention network (FOANet) which introduced a separate channel for every focus region (global, right/left hand) and modality (RGB, depth, and flow). Zhang *et al.* [68] proposed an attention mechanism embedding into the ConvLSTM network, including attention analysis in ConvLSTM gates for spatial global average pooling and fully connected operations. Li *et al.* [57] proposed a spatiotemporal attention-based ResC3D network to focus on the gesture itself. These attention-based methods have shown high performances for gesture recognition, as shown in Table V.

C. Multimodality Fusion Scheme

For the task of RGB-D gesture recognition, fusion mechanisms are widely considered [28], [32], [33], [35], [43], [51]. This kind of scheme consolidates the scores generated by networks that are fed with different modalities. Among these methods, the averaging [28], [33], [35], [43] and multiply [32], [51] score fusions are two of the most frequently applied. Li *et al.* [31], Zhu *et al.* [33], and Miao *et al.* [27] adopted feature-level fusion. The former methods [31], [33] directly blend the features of RGB and

depth modalities in a parallel or serial way, which simply average or concatenate. Considering the relationship between features from different modalities that share the same label, Miao *et al.* [27] adopted a statistical analysis-based fusion method—canonical correlation analysis, and Li *et al.* [31] adopted an extension version of discriminative correlation analysis, which tries to maximize the inner class pairwise correlations across modalities and intraclass differences within one feature set. Hu *et al.* [50] paid more attention to the fusion scheme and designed a new layer comprised of a group of networks called the adaptive hidden layer, which serves as a selector to weight features from different modalities of data. Lin *et al.* [52] developed an adaptive scheme for setting weights of each voting subclassifier via a fusion layer, which can be learned directly by the CNNs.

D. Other Techniques to Boost Performance

Multiple Modalities: Based on the available RGB and depth data modalities in the proposed datasets, additional data modalities have been considered by researchers. Li *et al.* [78] generated saliency maps to focus on image parts relevant to gesture recognition and used optical flow [31] to learn features from RGB image motion vectors. Wang *et al.* [49] and Asadi-Aghbolaghi *et al.* [61] extended optical flow from RGB videos to construct RGB and depth optical flow from RGB and depth videos, respectively. Another strategy uses the skeleton information as an extra modality via regional multiperson pose estimation (RMPE) [21], [69].

Data Augmentation: Data augmentation is another common way to boost performance. Miao *et al.* [27] focused on data augmentation to increase overall dataset size while Zhang *et al.* [44] mainly augmented data to balance the number of samples among different categories, including translation, rotation, the Gaussian smoothing, and contrast adjustment.

Pretrained Models: Some C3D-implemented methods [31], [33], [43] are pretrained on external datasets, such as sports-1M [70]. In terms of cross-modality finetuning, Zhu *et al.* [43] first trained the networks with RGB and depth data from scratch and then finetuned the depth one with the model trained from RGB data. The same process is done for the RGB model. The result of cross-modality finetuning [43] showed an improvement of 6% and 8% for RGB and depth inputs, respectively.

E. Summary of State-of-the-Art Methods

We summarize the previous methods features and compare them in Tables V and VI for isoGD and conGD gesture recognition datasets, respectively. All methods are published in the last three years. For isolated gesture recognition on the IsoGD dataset, Table V, all methods except [25] are based on deep learning. The recognition rate is improved by 58% from the 24.19% of the handcrafted method of [25] to the 82.17% of the 2-D CNN fusion strategy of [54]. For continuous gesture recognition on the ConGD dataset, Table VI, the performance has also been improved considerably for both MJI and CSR metrics since 2017.

F. Discussion

In this section, we review the techniques on both isolated and continuous gesture recognition based on RGB-D data. After the release of the large-scale IsoGD and ConGD datasets, new methods have pushed the development of gesture recognition algorithms. However, there are challenges faced by the available methods that allow us to outline several future research directions for the development of deep learning-based methods for gesture recognition.

Fusion of RGB-D Modalities: Most methods [28], [32], [33], [35], [43] considered RGB and depth modality as a separate channel and fused them at a later stage by concatenation or score voting, without fully exploiting the complementary properties of both visual modalities. Therefore, cooperative training using RGB-D data would be a promising and interesting research direction.

Attention-Based Mechanism: Some methods [34], [38], [54] used hand detectors to first detect hand regions and then designed different strategies to extract local and global features for gesture recognition. However, these attention-based methods need hard to train specialized detectors to find hand regions properly. It would be more reasonable to consider sequence modeling self-attention [71], [72] and exploit it for dynamic gesture recognition.

Simultaneous Gesture Segmentation and Recognition: The existing continuous gesture recognition works [28], [34], [40], [48] first detect the first and end points of each isolated gesture, and then train/test each segment separately. This procedure is not suitable for many real applications. Therefore, simultaneous gesture segmentation and recognition would be an interesting line to be explored.

Joint Structure Learning: A structure attention mechanism can be further explored. Some works [52], [54] train each attention part (i.e., arm, gesture) separately and fuse the scores of several waterworks to obtain the final recognition result. However, it cannot consider the structure information among the relationships of body parts. We believe gesture recognition will benefit from joint structure learning (i.e., body, hand, arm, and face).

Efficient and Fast Networks: We discussed different works that benefited from the fusion and combination of different trained models and modalities. However, these are highly complex strategies from a computational perspective. For real applications, lightweight networks would be preferred.

Besides, there is one criticism encountered from the challenges. In real gesture-based applications (i.e., sign language recognition), the interval of neighbor gestures sometimes would be not obvious. However, in the ConGD and CGD datasets, the presence of silence between neighbor gestures always happens. In future work, the release of the annotated continuous sign language datasets without nonobvious silences would push further the research in the field.

IV. Temporal Segmentation Benchmark

Here, we propose a benchmark method, namely, the Bi-LSTM network, for temporal segmentation. Before it, we

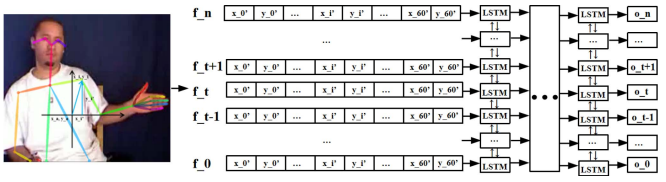


Fig. 3. Gray point $(\bar{x}, \bar{y})$ is determined by averaging all detected key points. (x_i, y_i) is the absolute coordinate of the ith key point, and (x_i', y_i') is the relative coordinate of the ith key point. We calculate the relative coordinates of key points in each frame and feed them to the Bi-LSTM network.

TABLE VII
TEMPORAL SEGMENTATION (CSR) METHODS COMPARISON ON VALIDATION AND TEST SETS ON THE CONGD DATASET

CSR* / Mehtod	Validation Set					Testing Set				
	0.5	0.6	0.7	0.8	0.9	0.5	0.6	0.7	0.8	0.9
Wang *et al.*[41]	0.857	0.7954	0.752	0.6997	0.5908	0.7711	0.6963	0.6636	0.6265	0.5497
Chai *et al.*[39]	-	-	-	-	-	0.709	0.5278	0.3213	0.1458	0.0499
Camgoz *et al.*[40]	-	-	-	-	-	0.7715	0.7008	0.6603	0.6054	0.5216
Liu *et al.*[35]	0.9313	0.9122	0.9034	0.8895	0.8132	0.9237	0.9032	0.8917	0.873	0.7750
Wang *et al.*[36]	0.857	0.7954	0.7520	0.6997	0.598	0.7711	0.6963	0.6636	0.6265	0.5497
pigou *et al.*[38]	0.7247	0.6625	0.6159	0.5634	0.4772	0.7313	0.6642	0.6241	0.5722	0.4951
Camgoz *et al.*[37]	0.8856	0.8421	0.8213	0.8024	0.7375	0.8833	0.8441	0.8254	0.8038	0.7187
Bi-LSTM (ours)	**0.9784**	**0.9699**	**0.9668**	**0.9638**	**0.9095**	**0.9765**	**0.9686**	**0.9639**	**0.9522**	**0.7876**

*column header 0.5 to 0.9 are IoU thresholds for CSR.

illustrate the drawbacks of the current temporal segmentation methods.

A. Drawbacks of Temporal Segmentation Methods

1) Handcrafted Hand Motion Extraction: Some methods [25], [28], [40], [48] first measure the quantity of movement (QoM) for each frame in a multigesture sequence and threshold the QoM to obtain candidate boundaries. Then, a sliding window is adopted to refine the candidate boundaries to produce the final boundaries of the segmented gesture sequences in a multigesture sequence. However, it captures not only hand motions but also the background movements that may be harmful to temporal segmentation.

2) Unstable Hand Detector: Some methods [34], [38] used the faster R-CNN [73] to build the hand detector. Due to the high degree of freedom of human hands, it is very hard to tackle some intractable environments, such as hand-self occlusion and drastically hand shape changing. The errors of hand detection would considerably reduce the performance of temporal segmentation.

3) Strong Prior Knowledge Requirement: Most previous methods (e.g., [25], [34], [38], [40], and [48]) use prior knowledge (e.g., a performer always raises hands to start a gesture and puts hands down after performing a gesture). The strong prior knowledge (i.e., the hand must lay down after performing another gesture) is not practical for real applications.

In contrast to the previous methods, we did not only use human hands but also the arm/body information [34], [38]. Moreover, we designed a Bi-LSTM segmentation network to determine the start–end frames of each gesture automatically without requiring specific prior knowledge.

B. Proposed Bi-LSTM Method

We treat the temporal segmentation as a binary classification problem. The flowchart is shown in Fig. 3. We first use the convolutional pose machine (CPM) algorithm[6] [74]–[76] to estimate the human pose, which consists of 60 keypoints (18 keypoints for human body and 21 keypoints for left and right hands, respectively). The keypoints are shown in the left part of Fig. 3. Therefore, the human gesture/body from an image is represented by these keypoints. For the t-th frame of a video, the gesture is represented by a 120-D (2×60) vector V_t in the following:

$$V_t = \{(x_i - \bar{x}, y_i - \bar{y}), i = 1, \ldots, 60\} \tag{4}$$

[6]https://github.com/CMU-Perceptual-Computing-Lab/openpose

where the coordinate of the ith keypoint is represented by (x_i, y_i), the average coordinate of all detected keypoints is denoted by $(\bar{x}, \bar{y})$, and $\bar{x} = (1/n_t') \sum_{k=1}^{n_t'} x_k$, $\bar{y} = (1/n_t') \sum_{k=1}^{n_t'} y_k$, n_t' is the number of detected keypoints of frame t.

We use the data $\{(V_t, g_t)|t = 1, \ldots, m\}$ to train the Bi-LSTM network [66], where m is the total number of frames in the video, and g_t is the start and end frames indicator of a gesture, that is, $g_t = 1$ indicates the start and last frames of a gesture, and $g_t = 0$ for other frames. The Bi-LSTM network combines the bidirectional RNNs (BRNNs) [77] and the LSTM, which captures long-range information in bidirections of inputs. The LSTM unit $\mathcal{H}$ is implemented by the following composite function:

$$\begin{aligned} i_t &= \sigma(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \\ f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \\ c_t &= f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \\ o_t &= \sigma(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \\ h_t &= o_t \tanh(c_t) \end{aligned} \tag{5}$$

where σ_t is the activation function, and i_t, f_t, c_t, o_t, and h_t are the input gate, forget gate, output gate, cell activation vector, and the hidden vector at time t, respectively. For the Bi-LSTM, the network computes both the forward and backward hidden vectors $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ at time t, and the output sequence y_t as

$$y_t = W_{\overrightarrow{h}_t y}\overrightarrow{h}_t + W_{\overleftarrow{h}_t y}\overleftarrow{h}_t + b_y. \tag{6}$$

We design four hidden layers in the Bi-LSTM network in Fig. 3. Notably, if a frame is within the segment of a gesture, we annotate it as the positive sample; otherwise, it is treated as negative. To mitigate the class imbalance issue, we assign different weights to the positive and negative samples. The objective function is defined as

$$J(\theta) = -\frac{1}{m}\left[\sum_{c=0}^{m-1}\sum_{i=0}^{1}\omega_i \log \frac{e^{\theta_i^T x_c}}{\sum_{j=0}^{k} e^{\theta_j^T x_c}}\right] \tag{7}$$

where θ is the parameter matrix of the softmax function and w_i is the weight used to mitigate the class imbalance issue. According to our statistics, the ratio of the positive and negative samples is approximately 1:40. Thus, we set $(w_0/w_1) = 40$ (w_0 is the weight penalty of positive samples) to balance the loss terms of positive and negative samples in the training phase.

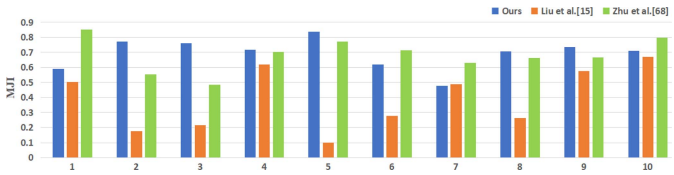


Fig. 4. Comparison of the first ten classes of the ConGD dataset for MJI metric. For most of the classes (six out of ten classes), our method achieves the best performance. From Table VI, the MJI of our method, Liu *et al.* [34] and Zhu *et al.* [58] for all classes are 0.7179, 0.6103, and 0.7163, respectively. Note that we only show MJIs of the first ten classes.

Fig. 5. Examples of failure cases by the CPM method [76] for keypoint detection, especially for hand keypoints. Gestue ID: (a) 1, (b) 6, (c) 7, and (d) 10.

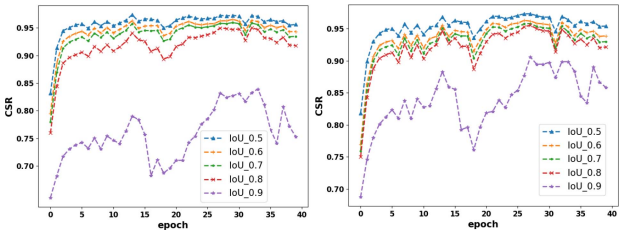


Fig. 6. CSR curve of the proposed Bi-LSTM method on the ConGD dataset. Up: test set. Bottom: validation set.

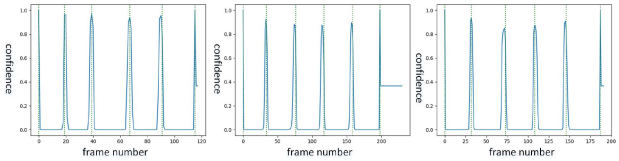


Fig. 7. Some results of the proposed Bi-LSTM method on three longest sequence videos of the ConGD dataset. Green point: ground truth of the segmentation point. Blue line is the confidence of the nonsegmentation point.

The gradient of the objective function is computed as

$$\nabla_{\theta_v} J(\theta) = -\frac{1}{m}\left[\sum_{c=0}^{m-1}\sum_{i=0}^{k}\omega_i x_c\left(\mathbb{I}_{[v=j]} - \frac{e^{\theta_i^T x_c}}{\sum_{j=0}^{k} e^{\theta_j^T x_c}}\right)\right] \quad (8)$$

where $\nabla_{\theta_v} J(\theta)$ is the gradient with respect to the parameter θ_v, and $\mathbb{I}_{[v=j]}$ is the indicator function, that is, $\mathbb{I}_{[v=j]} = 1$ if and only if $v = j$.

In this way, we use the learned model by the Bi-LSTM network to predict the probability of each frame whether it is the start or end frames. If the probability value of a frame is large than 0.5, this frame is treated as start or end frames.

V. Experiments

In this section, we evaluate and compare our proposed gesture recognition by the segmentation strategy on the ConGD dataset. First, the experimental setup is presented, including the running environments and settings. Then, the performances and comparisons on the ConGD dataset are given.

Our experiments are conducted on an NVIDIA Titan Xp GPU. The input of the Bi-LSTM network is a 120-D vector. We use the Adam algorithm to optimize the model with the batch size 120. The learning rate starts from 0.01 and the models are trained for up to 50 epochs.

The performance of the proposed Bi-LSTM method for temporal segmentation is shown in Table VII, which achieves 0.9668 and 0.9639 for CSR@IoU = 0.7 on both validation and testing sets of ConGD. After temporal segmentation, we use the model of [52] to perform final gesture recognition. The results are also shown in Table VI, where MJI = 0.6830 and 0.7179 on the validation and test sets, respectively. Based on MJI and CSR, our method achieves the best performance. Although the metric of MJI depends on both temporal segmentation and final classification, the recognition performance of MJI can still benefit from an accurate temporal segmentation, such as the proposed Bi-LSTM method.

We also provide comparisons for each category on the ConGD dataset in Fig. 4. Here, our method (overall MJI: 0.6830 for the validation set and 0.7179 for the test set) is compared with two state-of-the-art methods [34] (overall MJI: 0.5163 for the validation set and 0.6103 for the test set) and [58] (overall MJI: 0.5368 for the validation set and 0.7163 for the test set) for each category.

For illustration purposes, we show MJIs scores of the first ten classes in Fig. 4. It shows that our method achieves the best performance among competing methods for most classes (six out of ten classes). Some cases of failure of our approach are shown in Fig. 5. They are mainly produced by self-occlusions and hand keypoints detection failures, which lead to inaccurate inputs used by our models. As future work, more accurate hand pose detection strategies would enhance the overall performance of the proposed solution.

Fig. 6 shows the CSR curve in each epoch under different IoU thresholds from 0.5 to 0.9. One can see that when the IoU threshold is between 0.5 and 0.8, the CSR is very stable after three epochs. When IoU is equal to 0.9, the training epochs for the CSR increases. This is because the correct condition is more strict (> 90% overlapped region will be treated as the correct one) and it will cost more time to seek the best CSR. Our proposed Bi-LSTM method can obtain very stable results under different IoU thresholds. For example, even the IoU is equal to 0.9, the CSR of our method still is higher than 0.9. Alternative temporal segmentation methods [34], [36], [37], [40] are relatively inferior (the best is about 0.81 in [34]) on the validation set. Also, our Bi-LSTM can obtain the best performance on the test set of the ConGD dataset.

Then, we randomly select 1000 video sequences in the ConGD datasets to check for computational requirements. It required about 0.4 s under the GPU environment [*NVIDIA TITAN X (Pascal)*] and 6 s on the CPU environment [*Intel Core i7-5820K@3.30* GHz] for the proposed Bi-LSTM method. It demonstrates the proposed Bi-LSTM method is ultrahigh-speed processing (∼0.4 ms/video-GPU, ∼6 ms/video-CPU). We note that the processing time of the CPM algorithm is about 283 ms/f on the GPU environment.

Finally, we selected the three longest video sequences of the ConGD dataset, and the segmentation results of the proposed Bi-LSTM method are shown in Fig. 7. The green points are the ground truth of the segmentation point, while the blue line is the confidence of positive responses. These three videos have more than 100 frames and contain at least five gestures. Compared with the videos with a fewer number of gestures, the dense gestures make it hard to find the segment points accurately. However, our Bi-LSTM method can mark the start and end points of each gesture, and the segmentation for all the gestures are with confidence over 0.8.

VI. Conclusion

In this article, we proposed IsoGD and ConGD datasets for the task of isolated and continuous gesture recognition, respectively. Both datasets are the current largest datasets for dynamic gesture recognition. Based on both datasets, we have run challenges in ICPR 2016 and ICCV 2017 workshops, which attracted more than 200 teams around the world and pushed the state of the art for gesture recognition. Then, we reviewed the last 3-years methods for gesture recognition based on the provided datasets. Besides, we proposed the Bi-LSTM method for temporal segmentation. We expect the proposed datasets to push the research in gesture recognition.

References

[1] I. Guyon, V. Athitsos, P. Jangyodsuk, H. J. Escalante, and B. Hamner, "Results and analysis of the chalearn gesture challenge 2012," in *Advances in Depth Image Analysis and Applications*. Heidelberg, Germany: Springer, 2013, pp. 186–204.

[2] S. Escalera *et al.*, "Multi-modal gesture recognition challenge 2013: Dataset and results," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2013, pp. 445–452.

[3] S. Ruffieux, D. Lalanne, and E. Mugellini, "ChairGest: A challenge for multimodal mid-air gesture recognition for close HCI," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2013, pp. 483–488.

[4] L. Liu and L. Shao, "Learning discriminative representations from RGB-D video data," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 1493–1500.

[5] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural network," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4125–4207.

[6] C. Cao, Y. Zhang, Y. Wu, H. Lu, and J. Cheng, "Egocentric gesture recognition using recurrent 3D convolutional neural networks with spatiotemporal transformer modules," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3763–3771.

[7] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 1–10.

[8] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2010, pp. 9–14.

[9] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from RGBD images," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2012, pp. 842–849.

[10] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *Int. J. Robot. Res.*, vol. 32, no. 8, pp. 951–970, 2013.

[11] Q. De Smedt, H. Wannous, and J.-P. Vandeborre, "Skeleton-based dynamic hand gesture recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshops*, 2016, pp. 1–9.

[12] Q. de Smedt, H. Wannous, J.-P. Vandeborre, J. Guerry, and B. L. Saux, "SHREC'17 track: 3D hand gesture recognition using a depth and skeletal dataset," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1–6.

[13] B. Ni, G. Wang, and P. Moulin, "RGBD-HUDAACT: A color-depth video database for human daily activity recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2011, pp. 1147–1153.

[14] J. F. Hu, W. S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 5344–5352.

[15] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, 2011, pp. 2556–2563.

[16] K. Soomro, A. R. Zamir, and M. Shah. (2012). *UCF101: A Dataset of 101 Human Actions Classes From Videos in the Wild*. [Online]. Available: https://arxiv.org/abs/1212.0402

[17] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 6299–6308.

[18] J. Wan, Q. Ruan, W. Li, and S. Deng, "One-shot learning gesture recognition from RGB-D data using bag of features," *J. Mach. Learn. Res.*, vol. 14, no. 42, pp. 2549–2582, 2013.

[19] I. Lüsi, S. Escarela, and G. Anbarjafari, "Sase: RGB-Depth database for human head pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 325–336.

[20] H. J. Escalante *et al.*, "Chalearn joint contest on multimedia challenges beyond visual analysis: An overview," in *Proc. Int. Conf. Pattern Recognit.*, 2016, pp. 67–73.

[21] J. Wan *et al.*, "Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 3189–3197.

[22] O. Russakovsky *et al.*, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2014.

[23] S. Escalera *et al.*, "Chalearn looking at people challenge 2014: Dataset and results," in *Proc. ChaLearn LAP Workshop (ECCV)*, 2014, pp. 459–473.

[24] S. Escalera, V. Athitsos, and I. Guyon, "Challenges in multi-modal gesture recognition," in *Gesture Recognition*. Cham, Switzerland: Springer, 2017, pp. 1–60.

[25] J. Wan, Y. Zhao, S. Zhou, I. Guyon, S. Escalera, and S. Z. Li, "Chalearn looking at people RGB-D isolated and continuous datasets for gesture recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 56–64.

[26] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[27] Q. Miao *et al.*, "Multimodal gesture recognition based on the RESC3D network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3047–3055.

[28] H. Wang, P. Wang, Z. Song, and W. Li, "Large-scale multimodal gesture segmentation and recognition based on convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3138–3146.

[29] L. Zhang, G. Zhu, P. Shen, J. Song, S. A. Shah, and M. Bennamoun, "Learning spatiotemporal features using 3DCNN and convolutional LSTM for gesture recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3120–3128.

[30] J. Duan, J. Wan, S. Zhou, X. Guo, and S. Z. Li, "A unified framework for multi-modal isolated gesture recognition," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 14, no. 1s, pp. 1–16, 2018.

[31] Y. Li *et al.*, "Large-scale gesture recognition with a fusion of RGB-D data based on saliency theory and C3D model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2956–2964, Oct. 2018.

[32] P. Wang, W. Li, S. Liu, Z. Gao, C. Tang, and P. Ogunbona, "Large-scale isolated gesture recognition using convolutional neural networks," in *Proc. Int. Conf. Pattern Recognit.*, 2016, pp. 7–12.

[33] G. Zhu, L. Zhang, L. Mei, J. Shao, J. Song, and P. Shen, "Large-scale isolated gesture recognition using pyramidal 3D convolutional networks," in *Proc. Int. Conf. Pattern Recognit.*, 2016, pp. 19–24.

[34] Z. Liu, X. Chai, Z. Liu, and X. Chen, "Continuous gesture recognition with hand-oriented spatiotemporal feature," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3056–3064.

[35] H. Wang, P. Wang, Z. Song, and W. Li, "Large-scale multimodal gesture recognition using heterogeneous networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3129–3137.

[36] N. C. Camgoz, S. Hadfield, and R. Bowden, "Particle filter based probabilistic forced alignment for continuous gesture recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 3079–3085.

[37] L. Pigou, M. Van Herreweghe, and J. Dambre, "Gesture and sign language recognition with temporal residual networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3086–3093.

[38] X. Chai, Z. Liu, F. Yin, Z. Liu, and X. Chen, "Two streams recurrent neural networks for large-scale continuous gesture recognition," in *Proc. Int. Conf. Pattern Recognit.*, 2016, pp. 31–36.

[39] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "Using convolutional 3D neural networks for user-independent continuous gesture recognition," in *Proc. Int. Conf. Pattern Recognit.*, 2016, pp. 49–54.

[40] P. Wang, W. Li, S. Liu, Y. Zhang, Z. Gao, and P. Ogunbona, "Large-scale continuous gesture recognition using convolutional neural networks," in *Proc. Int. Conf. Pattern Recognit.*, 2016, pp. 13–18.

[41] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[42] K. Simonyan and A. Zisserman. (2014). *Very Deep Convolutional Networks for Large-Scale Image Recognition*. [Online]. Available: https://arxiv.org/abs/1409.1556

[43] G. Zhu, L. Zhang, P. Shen, and J. Song, "Multimodal gesture recognition using 3D convolution and convolutional LSTM," *IEEE Access*, vol. 5, pp. 4517–4524, 2017.

[44] Z. Zhang, S. Wei, Y. Song, and Y. Zhang, "Gesture recognition using enhanced depth motion map and static pose map," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 238–244.

[45] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.

[46] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[47] X. Liu *et al.*, "Late fusion incomplete multi-view clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2410–2423, Oct. 2019.

[48] J. Wan, G. Guo, and S. Z. Li, "Explore efficient local features from RGB-D data for one-shot learning gesture recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1626–1639, Aug. 2016.

[49] P. Wang, W. Li, Z. Gao, Y. Zhang, C. Tang, and P. Ogunbona, "Scene flow to action map: A new representation for RGB-D based action recognition with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 595–604.

[50] T.-K. Hu, Y.-Y. Lin, and P.-C. Hsiu, "Learning adaptive hidden layers for mobile gesture recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 6934–6942.

[51] P. Wang, W. Li, J. Wan, P. Ogunbona, and X. Liu, "Cooperative training of deep aggregation networks for RGB-D action recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[52] C. Lin, J. Wan, Y. Liang, and S. Z. Li, "Large-scale isolated gesture recognition using a refined fused model based on masked RES-C3D network and skeleton LSTM," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, 2018, pp. 52–58.

[53] P. Wang, W. Li, Z. Gao, C. Tang, and P. O. Ogunbona, "Depth pooling based large-scale 3-D action recognition with convolutional neural networks," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1051–1061, May 2018.

[54] P. Narayana, J. R. Beveridge, and B. A. Draper, "Gesture recognition: Focus on the hands," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 5235–5244.

[55] C. Zhu, J. Yang, Z. Shao, and C. Liu, "Vision based hand gesture recognition using 3D shape context," *IEEE/CAA J. Automatica Sinica*, early access, May 27, 2019, doi: 10.1109/JAS.2019.1911534.

[56] G. Zhu *et al.*, "Redundancy and attention in convolutional LSTM for gesture recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 4, pp. 1323–1335, Apr. 2020.

[57] Y. Li, Q. Miao, X. Qi, Z. Ma, and W. Ouyang, "A spatiotemporal attention-based RESC3D model for large-scale gesture recognition," *Mach. Vis. Appl.*, vol. 30, no. 5, pp. 875–888, 2019.

[58] G. Zhu, L. Zhang, P. Shen, J. Song, S. A. A. Shah, and M. Bennamoun, "Continuous gesture segmentation and recognition using 3DCNN and convolutional LSTM," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 1011–1021, Apr. 2019.

[59] N. N. Hoang, G.-S. Lee, S.-H. Kim, and H.-J. Yang, "Continuous hand gesture spotting and classification using 3D finger joints information," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2019, pp. 539–543.

[60] E. H. Land and J. J. McCann, "Lightness and retinex theory," *J. Opt. Soc. America*, vol. 61, no. 1, pp. 1–11, 1971.

[61] M. Asadi-Aghbolaghi, H. Bertiche, V. Roig, S. Kasaei, and S. Escalera, "Action recognition from RGB-D data: Comparison and fusion of spatio-temporal handcrafted features and deep strategies," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3179–3188.

[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.

[63] J. Nagi *et al.*, "Max-pooling convolutional neural networks for vision-based hand gesture recognition," in *Proc. IEEE Int. Conf. Signal Image Process. Appl.*, 2011, pp. 342–347.

[64] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 773–787, Apr. 2017.

[65] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.

[66] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional LSTM and other neural network architectures," *Neural Netw.*, vol. 18, nos. 5–6, pp. 602–610, 2005.

[67] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.

[68] Y. Zhang, C. Cao, J. Cheng, and H. Lu, "EgoGesture: A new dataset and benchmark for egocentric hand gesture recognition," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1038–1050, May 2018.

[69] H.-S. Fang, S. Xie, Y.-W. Tai, and C. Lu, "Rmpe: Regional multi-person pose estimation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2334–2343.

[70] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2014, pp. 1725–1732.

[71] T. Shen, T. Zhou, G. Long, J. Jiang, S. Wang, and C. Zhang. (2018). *Reinforced Self-Attention Network: A Hybrid of Hard and Soft Attention for Sequence Modeling*. [Online]. Available: https://arxiv.org/abs/1801.10296

[72] T. Shen, T. Zhou, G. Long, J. Jiang, and C. Zhang. (2018). *Bi-Directional Block Self-Attention for Fast and Memory-Efficient Sequence Modeling*. [Online]. Available: https://arxiv.org/abs/1804.00857

[73] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

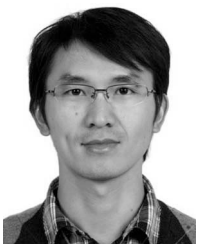
[74] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 7291–7299.

[75] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 1145–1153.

[76] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 4724–4732.

[77] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, Nov. 1997.

[78] Y. Li *et al.*, "Large-scale gesture recognition with a fusion of RGB-D data based on optical flow and the C3D model," *Pattern Recognit. Lett.*, vol. 119, pp. 187–194, 2019.



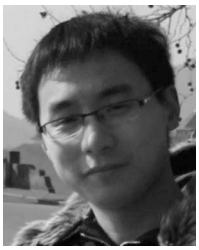
Jun Wan (Senior Member, IEEE) received the B.S. degree from the China University of Geosciences, Beijing, China, in 2008, and the Ph.D. degree from Beijing Jiaotong University, Beijing, in 2015.

Since January 2015, he has been a Faculty Member with the Institute of Automation, Chinese Academy of Sciences, Beijing, where he currently serves as an Associate Professor. His main research interests include computer vision and machine learning.

Dr. Wan is an Associate Editor of *IET Biometrics*. He has served as a Co-Editor of special issues in IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE.

Chi Lin (Member, IEEE) received the B.S. degree (First Class Hons.) from the Faculty of Information Technology, Macau University of Science and Technology, Macau, China, in 2017. He is currently pursuing the M.S. degree with the University of Southern California, Los Angeles, CA, USA.

From 2015 to 2016, he was selected to participate in "Stars of Tomorrow Internship Program" in Microsoft Research Asia (MSRA). His research interests include machine learning, computer vision, and gesture recognition.



Longyin Wen received the B.Eng. degree in automation from the University of Electronic Science and Technology of China, Chengdu, China, in 2010, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2015.

He is a Staff Scientist with the AI Lab, JD Finance, Mountain View, CA, USA. After that, he moved to the University at Albany, State University of New York, Albany, NY, USA, for postdoctoral research. From 2016 to 2018, he was a Computer Vision Scientist with GE Global Research, Niskayuna, NY, USA.

Yunan Li received the Ph.D. degree from the School of Computer Science and Technology, Xidian University, Xi'an, China, in 2019.

He currently serves as a Lecturer with Xidian University. His research interests include machine learning, computer vision, and gesture recognition.

Qiguang Miao (Senior Member, IEEE) received the Doctoral degree in computer application technology from Xidian University, Xi'an, China, in 2005.

He is a Professor and the Ph.D. Student Supervisor with the School of Computer Science and Technology, Xidian University. He has published over 100 papers in the significant domestic and international journals or conferences. His research interests include machine learning, intelligent image processing, and malware behavior analysis and understanding.

Sergio Escalera (Member, IEEE) received the Ph.D. degree in multiclass visual categorization systems from the Computer Vision Center, University of Alabama at Birmingham (UAB), Birmingham, AL, USA, in 2008.

He is an Associate Professor with the Department of Mathematics and Informatics, Universitat de Barcelona, Barcelona, Spain. He is an Adjunct Professor with the Universitat Oberta de Catalunya, Barcelona, Aalborg University, Aalborg, Denmark, and Dalhousie University, Halifax, NS, Canada.

Dr. Escalera obtained the 2008 Best Thesis Award on Computer Science at Universitat Autónoma de Barcelona. He leads the Human Pose Recovery and Behavior Analysis Group at UB and CVC. He is also a member of the Computer Vision Center at Campus UAB.

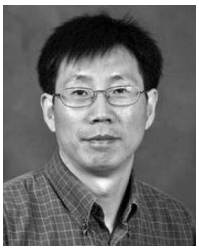
Gholamreza Anbarjafari (Senior Member, IEEE) received the Ph.D. degree in electrical and electronic engineering (major in image processing) from Eastern Mediterranean University, Famagusta, Cyprus, in 2010.

He is currently Director and Chief Data Scientist at PwC Advisory in Finland. He is also with Faculty of Engineering at Hasan Kalyoncu University, Turkey. He is the Founder of the Intelligent Computer Vision Lab, University of Tartu, Tartu, Estonia. He was also the Deputy Scientific Coordinator of the European Network on Integrating Vision and Language (iV&L Net) ICT COST Action IC1307. He has been a representative of Estonia for several other COST Actions.

Dr. Anbarjafari is an Associate Editor and a Guest Lead Editor of several journals, Special Issues and Book projects. He is the Chair of Signal Processing/Circuits and Systems/Solid-State Circuits Joint Societies Chapter of IEEE Estonian section.

Isabelle Guyon received the Ph.D. degree in physical sciences from University Pierre, Paris, France, in 1988.

She is a Chaired Professor of big data with the Universite Paris-Saclay (UPSud/INRIA), Saint-Aubin, France, specialized in statistical data analysis, pattern recognition, and machine learning. Prior to joining Paris-Saclay, she worked as an Independent Consultant and was a Researcher with AT&T Bell Laboratories, Murray Hill, NJ, USA. Her areas of expertise include computer vision and bioinformatics. Her recent interest is in applications of machine learning to the discovery of causal relationships.



Guodong Guo (Senior Member, IEEE) received the B.E. degree in automation from Tsinghua University, Beijing, China, the first Ph.D. degree in pattern recognition and intelligent control from the Chinese Academy of Sciences, Beijing, China, and the second Ph.D. degree in computer science from the University of Wisconsin–Madison, Madison, WI, USA, in 2006.

In 2018, he joined with Baidu Research, Beijing. He is an Associate Professor with the Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV, USA.

Stan Z. Li (Fellow, IEEE) received the B.Eng. degree from Hunan University, Changsha, China, in 1982, the M.Eng. degree from the National University of Defense Technology, Changsha, in 1985, and the Ph.D. degree from Surrey University, Guildford, U.K., in 1991.

He is a Chair Professor of artificial intelligence with Westlake University, Hangzhou, China. He was a Researcher and the Director of the Center for Biometrics and Security Research with Institute of Automation, Chinese Academy of Sciences, Beijing, China. He was a Researcher with Microsoft Research Asia, Beijing, and an Associate Professor with Nanyang Technological University, Singapore. He has published over 500 papers with Google scholar index of over 42 000 and h-index of 95.