IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 8, 2022, accepted January 24, 2022, date of publication February 14, 2022, date of current version February 22, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3151362

Using Motion History Images With 3D Convolutional Networks in Isolated Sign Language Recognition

OZGE MERCANOGLU SINCAN[1], (Member, IEEE), AND HACER YALIM KELES[2]
[1]Computer Engineering Department, Ankara University, 06830 Ankara, Turkey
[2]Computer Engineering Department, Hacettepe University, 06800 Ankara, Turkey

Corresponding author: Ozge Mercanoglu Sincan (omercanoglu@ankara.edu.tr)

This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grant 217E022.

ABSTRACT Sign language recognition using computational models is a challenging problem that requires simultaneous spatio-temporal modeling of the multiple sources, i.e. faces, hands, body, etc. In this paper, we propose an isolated sign language recognition model based on a model trained using Motion History Images (MHI) that are generated from RGB video frames. RGB-MHI images represent spatio-temporal summary of each sign video effectively in a single RGB image. We propose two different approaches using this RGB-MHI model. In the first approach, we use the RGB-MHI model as a motion-based spatial attention module integrated into a 3D-CNN architecture. In the second approach, we use RGB-MHI model features directly with the features of a 3D-CNN model using a late fusion technique. We perform extensive experiments on two recently released large-scale isolated sign language datasets, namely AUTSL and BosphorusSign22k. Our experiments show that our models, which use only RGB data, can compete with the state-of-the-art models in the literature that use multi-modal data.

INDEX TERMS 3D-CNN, attention, deep learning, motion history image, sign language recognition.

I. INTRODUCTION

According to the World Federation of the Deaf (WFD), there are 70 million deaf people around the world [1]. People in deaf communities use sign language for communication with each other and with hearing people. Sign languages are visual languages that use manual articulations of hands in combination with facial expression and body posture to represent signs. Each country has its own sign language, and sign languages have their own lexicon and grammar. The fact that the majority of hearing people do not know sign language poses a challenge for deaf communities in daily life. To overcome this challenge, computer vision researchers carry out studies on automatic Sign Language Recognition (SLR) systems.

In the literature, there are two main tracks of related research on the SLR domain; isolated and continuous SLR. While a video sample includes only one sign in an isolated SLR setting, continuous SLR videos contain multiple signs and require recognition of a sequence of multiple signs. Although continuous SLR is more challenging, isolated SLR is still an active research area. Particularly, the recently released new large-scale datasets provide unique challenges to the community [2]–[7]. According to the survey in [8], there is an exponential growth in isolated SLR studies between 1983-2020.

Computationally, isolated SLR can be considered similar to an action recognition problem. However, SLR involves the discrimination of a sequence of fine-grained local motions that are usually performed quickly. For similarly performed signs, only small differences make signs different; for instance, for some pairs of signs, hand gestures look very similar, yet there is a difference in the facial expressions. For some pairs of signs, hand motion trajectories look very similar, yet local hand gestures look slightly different. For some pairs of signs, although hand gestures and trajectories are the same, the number of repetitions of a gesture is different. To deal with these challenges, most studies

The associate editor coordinating the review of this manuscript and approving it for publication was Li He.

explicitly segment the hands or face regions in video frames, or use multiple modalities, such as depth, skeleton, etc. to obtain higher accuracies in sign classification.

In this research, we aim to propose an effective solution, which uses only RGB images in the sign videos without requiring any explicit part segmentation or additional data modalities. We create RGB-Motion History Images (RGB-MHI) for each video and train an SLR model using only RGB-MHI images. Based on this model, we propose two different approaches that use RGB-MHI images and RGB frames of sign videos: In the first approach, we experiment with a 3D Convolutional Neural Network (3D-CNN) using our RGB-MHI model features as an attention mechanism to focus on relevant parts of the video frames. In the second approach, RGB frames and RGB-MHI images are used together with a late fusion technique. Our empirical observations show that using RGB-MHI images with 3D-CNN models provides comparable performances with the state-of-the-art models that use multiple input modalities, such as depth, skeleton, hand and face segmentation, etc., or multiple model ensembles. The proposed solutions in this research are considerably cheaper with respect to their memory and computation requirements.

The remainder of the paper is organized as follows: In Section II, we summarize the related works. In Section III, we describe our proposed methods. In Section IV, we provide the experiments and model training details. Finally, we conclude the paper in Section V.

II. RELATED WORKS

Feature extraction is a crucial step for isolated sign language recognition as it is for other problems in computer vision. In the literature, some early studies utilized mainly hand features for sign language recognition by utilizing glove-based models, or using different devices, such as Leap Motion Controller (LMC), which aims to detect and track hands and fingers [9], [10]. Tracking only the hand region is not sufficient for general-purpose sign recognition. The signs in sign languages are identified by manual sources and non-manual sources simultaneously. While manual sources consist of 4 components: hand shape (finger configurations), orientation of the palm, movement of the hand, and position of the hand; non-manual sources include body posture and facial expressions such as head tilting, mouthing, etc. [11]. A powerful SLR system should follow all these body parts simultaneously. Therefore, in most recent SLR studies, the methods approach the problem holistically; analyzing all components of the signer data as a whole.

The early research in the field use hand-crafted features such as HOG, Hu moments, motion velocity vector, relative hand positioning [12]–[14] as features. After the release of the new large-scale datasets and advancements in deep neural networks, spatio-temporal sign features are successfully learned from the data distribution. Convolutional deep models are useful in this regard to learn spatial features automatically from data [15]–[17]. In these studies, after spatial features are extracted by 2D-CNNs, temporal information is modelled by a recurrent neural network. On the other hand, some studies use the combination of deep learning and traditional methods. Rastgoo *et al.* [18] used some hand-crafted features and 2D-CNNs to obtain spatial information. Then, they fuse all the features and feed them to LSTM for temporal feature extraction.

Recently, 3D-CNNs show remarkable performances in modeling the spatio-temporal patterns simultaneously in video frames. Joze and Koller [2] released a large-scale American Sign Language (ASL) dataset, namely MS-ASL, and provided baseline methods including 2D-CNN-LSTM and 3D-CNN based models. A 3D-CNN based I3D model [19] achieved the best result with a large improvement to the 2D-CNN based model. Li *et al.* [4] released another large-scale isolated ASL dataset namely WLASL, and they experimented with several appearance-based and pose-based deep learning methods; (a) 2D-CNN network and Gated Recurrent Unit (GRU), (b) 3D-CNN network, (c) human pose-based GRU (d) human pose-based temporal graph convolution network. Their results also show that 3D-CNN based I3D achieved the best results. Deep learning architectures that work with a sequence of images require considerable processing power and memory requirements. Therefore, some studies work on more efficient representations; Dos Santos *et al.* [20] created colored motion history images (MHI), namely the star RGBs, to represent video sequences. They trained two ResNets and combine these models with a soft-attention mechanism. They achieved significant accuracy rates in three different public datasets using only RGB data. Imran and Raman [21] also represented a video in a single image. They created three different kinds of motion templates: MHI, RGB motion image, dynamic image. In our proposed approach, we use a similar single MHI image representation to [20], yet we compute MHI images as in [22].

In SLR research, it is a common approach to use different types of data modalities such as depth, skeleton, optical flow, etc. in addition to RGB data, and combine these multiple modalities with a fusion technique in order to boost the accuracy rate. Jiang *et al.* [23], the winner team of the ChaLearn 2021 Looking at People Large Scale Signer Independent Isolated SLR CVPR Challenge [24], utilized different types of data modalities, e.g., RGB, skeleton, optical flow, depth, depth HHA, depth flow. They mention that multi-modal ensembles obtain a higher accuracy rate than every single modality. Some studies try to segment hand or face regions to use them as an additional modality. Luqman and El-Alfy [17] utilized Microsoft Kinect for this purpose. Firstly, they trained several 2D-CNN-LSTM models using color, depth, or optical flow data separately. Then, they used animation units of the signer face provided by the Kinect as input and trained a stacked LSTM model. They fused these manual and non-manual features at the classification level with their best 2D-CNN-LSTM model and improved the accuracy rate by about 3%. Some studies try to segment these

regions from RGB data without using any hardware but with some extra preprocessing. Using a pose estimation algorithm, such as OpenPose [25], is one of the popular methods to detect body, face, and hand keypoints. In [26] and [27], hand regions were cropped by creating a bounding box around the hand keypoints obtained using OpenPose. They trained several models separately using different types of data including cropped hands and proposed an ensemble model in order to increase the recognition accuracy. Gökçe *et al.* [28] also utilized OpenPose to crop both face and hand regions, and used these modalities together with the full-body images. Rastgoo *et al.* [29] used Faster R-CNN model [30] to detect hand regions, and they created three forms of input; original image, cropped hand image, and noisy cropped hand image. They proposed a Restricted Boltzmann Machine (RBM) using RGB and depth modalities to perform automatic hand sign language recognition.

In parallel to these, using attention mechanisms to focus on relevant spatial regions or temporal video frames without needing explicit segmentation have been also studied in the SLR domain. Shi *et al.* [31] proposed an iterative visual attention mechanism for ASL fingerspelling. They aimed to focus on signing hand with an attention model, which is based on a convolutional recurrent architecture. Their attention model reduced the area of interested regions iteratively, and while doing this, increased the resolution. Huang *et al.* [3] proposed an attention-based 3D-CNN for isolated SLR. In their proposed model, they extracted spatio-temporal features with a C3D model. A spatial attention mechanism was incorporated into C3D by using skeleton information of hands and arms. Then, they built a temporal attention-based model for classification. In our previous work [5], we proposed a baseline model for a new large-scale isolated Turkish Sign Language (AUTSL) dataset. We integrated a Feature Pooling Module and a temporal attention model to focus on more relevant spatio-temporal parts of the videos. Hu *et al.* [7] collected non-manual features aware isolated CSL dataset, which contains visually similar confusing signs that only differ in their non-manual features. They proposed 3D-Resnet-50 based [32] model that consists of two cooperative global and local enhancement modules to differentiate confusing signs. Global enhancement module contains a self-attention mechanism based on a Non-local Neural Network [33], with the aim of enhancing the global contextual relationship. In our proposed work, we utilize the RGB-MHI images as an attention model to assist our 3D-CNN model to focus on spatially salient regions.

III. PROPOSED METHOD

In this paper, we aim to construct a sign language recognition model using only RGB data, without using any explicit part segmentation of hands or face regions. In this regard, we create an RGB-MHI image that summarizes the entire video in a single video frame and propose a RGB-MHI model that learns a representation of relevant spatial and motion patterns using these single images. Then, we propose two different approaches utilizing this model. In the first approach, we use the RGB-MHI model as a motion-based attention mechanism to focus on relevant spatial regions. In the second one, we propose a fusion model that combines RGB and RGB-MHI features.

A. RGB-MOTION HISTORY IMAGE MODEL

Motion History Image (MHI) [34] is a static gray-scale image where more recently moving pixels are brighter; there are different variants of MHI image generation [22], [35]. In [22], MHI is estimated by the sum of the absolute values of differences between consecutive frames as in (1), where N is the number of frames in a video; I_t is the t^{th} video frame; (i, j) are the coordinates of a pixel in a frame and W represents the weight for the absolute difference, which is calculated as $W = t/N$. In [20], a modified version of (1) is proposed by including magnitude and phase information as well. They also propose an approach to represent MHIs as RGB images. In this approach, the frames in a video stream are split into three equal parts, and motion histories are calculated for each part separately. Then, a 3-channel colored MHI image, which is referred to as star RGB, is created. In this representation B-channel contains the motion history information from the first temporal region, the G-channel has the motion history information from the central one, and the R-channel from the last part. Dividing the video frames into 3 parts helps prevent temporal information loss.

$$M(i,j) = \sum_{t=2}^{N} |I_{t\text{-}1}(i,j) - I_t(i,j)|W \tag{1}$$

In our proposed method, we also consider using colored MHI images for SLR as in [20], yet while generating our RGB-MHI images we use the differences between two consecutive frames as in (1) for its simplicity. We depicted our colored RGB-MHI image generation approach in Fig. 1.

Our RGB-MHI model is based on ResNet-50 [36], which achieved successful results with star RGB in gesture recognition problem [20]. We use pre-trained ResNet-50 model on ImageNet [37] dataset, and fine-tune it to the used sign language datasets. However, RGB-MHI images are contextually different from normal images. Therefore, we made some ablation studies to select ResNet layers depending on their performances; we wanted to reduce the number of layers without sacrificing the model performance. However, we observed that the classification accuracies were lower than we expected when we used the Global Average Pooling (GAP) as it was used in the original ResNet model. We believe that losing the details in the spatial context affects model performances considerably with MHI images. Therefore, we removed the GAP layer and included an additional fully connected (FC) layer with 1024 neurons, with a ReLU non-linearity; the last FC layer follows this layer with a softmax function. We obtained the best results when we included all the convolutional layers in the

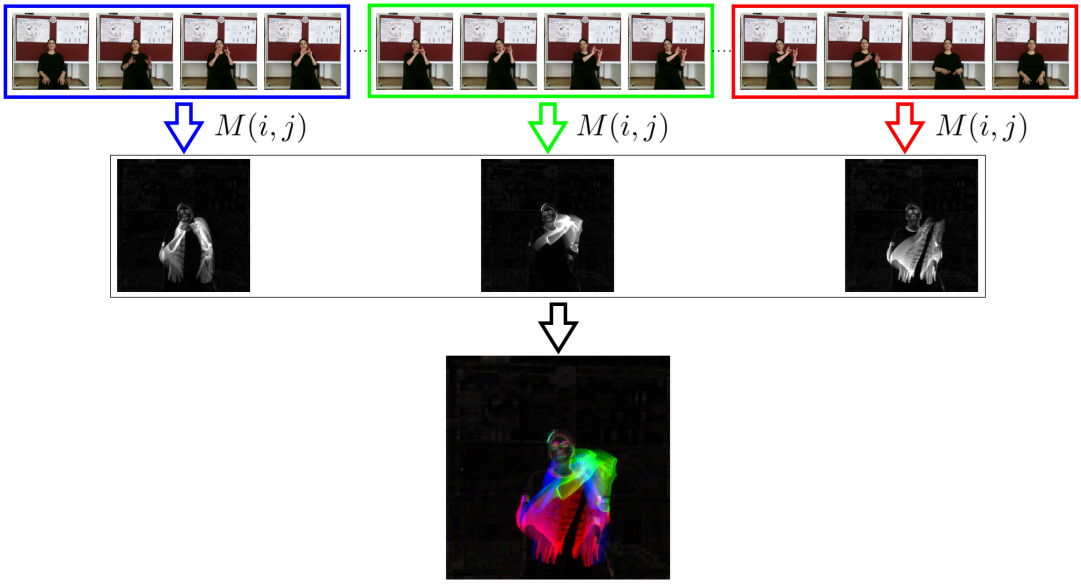

FIGURE 1. An example of creating an RGB-MHI image.

pre-trained ResNet-50 model. The details of the experimental results are provided in Section IV-C. As we expected, 2D-CNN models converge much faster when we use single RGB-MHI images than using a sequence of RGB video frames.

B. 3D-CNN MODELS

3D-CNN models have recently been widely used in video classification problems in different computer vision domains and provide state-of-the-art results. Motivated from this, in this study, we planned preliminary experiments with several successful 3D-CNN models to select the base 3D model that we can use together with our RGB-MHI model. In this context, we first worked with the C3D [38] model that is pre-trained on Sports-1M [39] dataset. Secondly, we experimented with the Inflated 3D ConvNet (I3D) [19] model that is pre-trained on Kinetics-400 [40] dataset. With the original C3D model, which has 8 convolutional layers, 5 pooling layers, 2 FC layers, we had a vanishing gradient problem. Therefore, we made some modifications in the architecture; we replaced the 2 FC layers with a global average pooling, and also we inserted a batch normalization (BN) layer after all the convolutional layers. Compared to the C3D model, the I3D model has fewer parameters and achieved better results; recently, it has been used on new sign language datasets, e.g., WLASL [4], MS-ASL [2], and BSL-1K [41]. In [41], it was observed that I3D achieved higher accuracy on two ASL datasets when pre-trained on a new large-scale British Sign Language dataset, BSL-1K, compared to Kinetics-400. Therefore, we experimented with an I3D model that is pre-trained using both Kinetics-400 and BSL-1K datasets.

In recent years, for efficiency reasons, 3D convolutions are separated into a 2D spatial convolution and a 1D temporal convolution. In these methods, 3D-CNNs are replaced with spatial and temporal separable 3D convolutions, i.e., $t \times k \times k$ is replaced with $1 \times k \times k$ followed by $t \times 1 \times 1$, where t is the width of the filter in time, and k is the width of the filter in spatial extent [42], [43]. In this work, we also experimented with two (2+1)D models, Separable 3D-CNN (S3D) [42] and R(2+1)D-18 [43]. S3D is the separable version of the I3D architecture and has much fewer parameters than the original I3D. On the other hand, although R(2+1)D does not reduce the number of parameters much, it eases the optimization process.

For all the architectures, we used publicly available pre-trained models and fine-tune them to the used sign language datasets. At the end of the preliminary experiments, we decided to work with the I3D model [19] that is pre-trained on BSL-1K [41].

C. 3D-CNN MODELS WITH ATTENTION MECHANISM

Attention mechanisms have been included in deep models in various ways to improve the performances of different tasks in computer vision and natural language processing domains. It has also been used in video classification problems, such as action recognition and sign language recognition, to focus on more relevant spatial or temporal parts of the stream and construct more robust models. In this study, we propose to utilize our RGB-MHI model as an attention model with the I3D model to focus on the spatial regions where there is a motion cue. In addition to our RGB-MHI based attention model, we also implemented a separate self-attention integrated I3D model. In this section, we first present the details of our self-attention integrated I3D model, then we describe our proposed motion-based attention mechanism.

1) SELF-ATTENTION MECHANISM

The convolutional operators in CNNs, process a region, i.e. computes the responses of filters, considering the values in their local neighborhood; hence, long-range dependencies are not captured. A self-attention mechanism aims to learn long-range dependencies [44]. The embedded Gaussian version of Non-Local (NL) blocks [33] is a kind of self-attention that is proposed for computer vision tasks. In this architecture, a non-local operation computes a response for an output position by the weighted sum of the features from all positions.

Motivated from [33], we incorporate NL blocks to some of the middle-layers of I3D architecture. In this context, we feed a feature map, $X \in R^{C\times T\times H\times W}$, into three different convolutional layers (θ, ϕ, g) with filters of size $1 \times 1 \times 1$, where C, T, H, W represent the channel, temporal, height, and width dimensions, respectively. The NL operation is described as in (2), where i is the index of output position, j is the index of all possible positions, X is the input feature map, and x is the output feature map. With all convolutional layers in (2), we reduce the channel dimension to half, thus the new channel dimension becomes $C/2$. Also, we apply a max pooling layer after ϕ and g to reduce the amount of pairwise computation. Then, the obtained feature map, x_{i}, is fed into another $1 \times 1 \times 1$ convolutional layer with C channel dimension, and the result is added to initial feature map X_{i} as in (3).

$$x_{\mathrm{i}} = softmax(\theta(X_{\mathrm{i}})^{\mathrm{T}}\phi(X_{\mathrm{j}}))g(X_{\mathrm{j}}) \tag{2}$$

$$z_{\mathrm{i}} = W_{\mathrm{z}}x_{\mathrm{i}} + X_{\mathrm{i}} \tag{3}$$

We conduct some experiments by adding a various number of NL blocks to different middle-layers. We achieve the best accuracy rate when we add 1 NL block after the third block (3a) of the I3D model. Our experimental results are given in the Table 3.

2) THE PROPOSED MOTION-BASED ATTENTION MECHANISM

In order to assess the salient regions using our pre-trained RGB-MHI model, we first apply channel-based global average pooling to the feature maps at the output of one of the middle layers. Then, we create an attention weight by passing the obtained feature matrix through softmax function and then normalizing the result between (0, 1) as follows:

$$\alpha = softmax(\sum_{j=1}^{W}\sum_{i=1}^{H}\frac{1}{C}\sum_{c=1}^{C}Y_{\mathrm{c,i,j}}) \tag{4}$$

$$\alpha_{norm} = \frac{\alpha - min(\alpha)}{max(\alpha) - min(\alpha)} \tag{5}$$

where $Y \in R^{C\times H\times W}$ is the feature map of a middle-layer in RGB-MHI model; $\alpha \in R^{H\times W}$; C, H, W represent the channel, height, and width dimensions, respectively. The normalized attention weights, α_{norm}, indicate the salient regions in a video. As ablation studies, we generated attention weights from three different layers of the RGB-MHI model, i.e. conv_2, conv_3, or conv_4. After the attention weights are obtained, it is element-wise multiplied with the feature maps in the I3D model, which has the same spatial size with α_{norm}; i.e. 2a, 3a, or 4a layers of the I3D.

Our proposed method is illustrated in Fig. 2a. In the Figure, we depict the motion-based attention aggregation in 3rd layer, since we observed the best accuracy rates using this layer.

D. FUSION OF I3D AND RGB-MHI MODELS

In the literature, some studies show that fusing multi-modalities or multi-model ensembles achieve higher accuracies than a single model. Therefore, we fuse our I3D and RGB-MHI models with a late fusion technique. In this approach, we utilize separate versions of I3D and RGB-MHI, without parameter updating. We assign w_1 and w_2 weights for the output of the last layers before softmax, and then calculate their weighted sum as in (6) for the final prediction. Our proposed fusion method is illustrated in Fig. 2b.

$$prediction = softmax(w_1x_1 + w_2x_2) \tag{6}$$

IV. EXPERIMENTS

This section describes the datasets, training details, various ablation studies, and comparisons to the state-of-the-art model performances in the literature.

A. DATASETS AND PREPROCESSING

We evaluate our proposed models on two very recently shared large-scale isolated sign language datasets: AUTSL [5] and BosphorusSign22k [6].

AUTSL [5] is a large-scale, signer independent, isolated TSL dataset that contains 226 signs, and 36,302 video samples. The dataset contains 43 signers; 31 of them are included in the training set, 6 are in the validation set, and the remaining 6 are in the test set. This benchmark provides a signer independent evaluation of the models. The methods are also challenged with 20 different backgrounds with dynamic elements, e.g., moving trees or moving people behind the signer.

BosphorusSign22k [6] is another large-scale, isolated TSL dataset that contains 744 signs, 22,542 video samples in which signs belong to health and finance domains, and also cover frequently used signs in daily activities. The dataset contains 6 signers; 1 of them is reserved for testing. In our experiments, we use 1 signer for the validation set. All signers are deaf people in this dataset.

1) DATA PREPROCESSING

Firstly, we apply pre-trained Faster R-CNN model [30] with ResNet-50 Feature Pyramid Network (FPN) backbone to the first frames of the videos in order to detect signer and eliminate some of the irrelevant background details. In some videos of AUTSL, some people are passing by behind the signer. Therefore, when there is more than one person in the video we choose the largest person bounding box since the

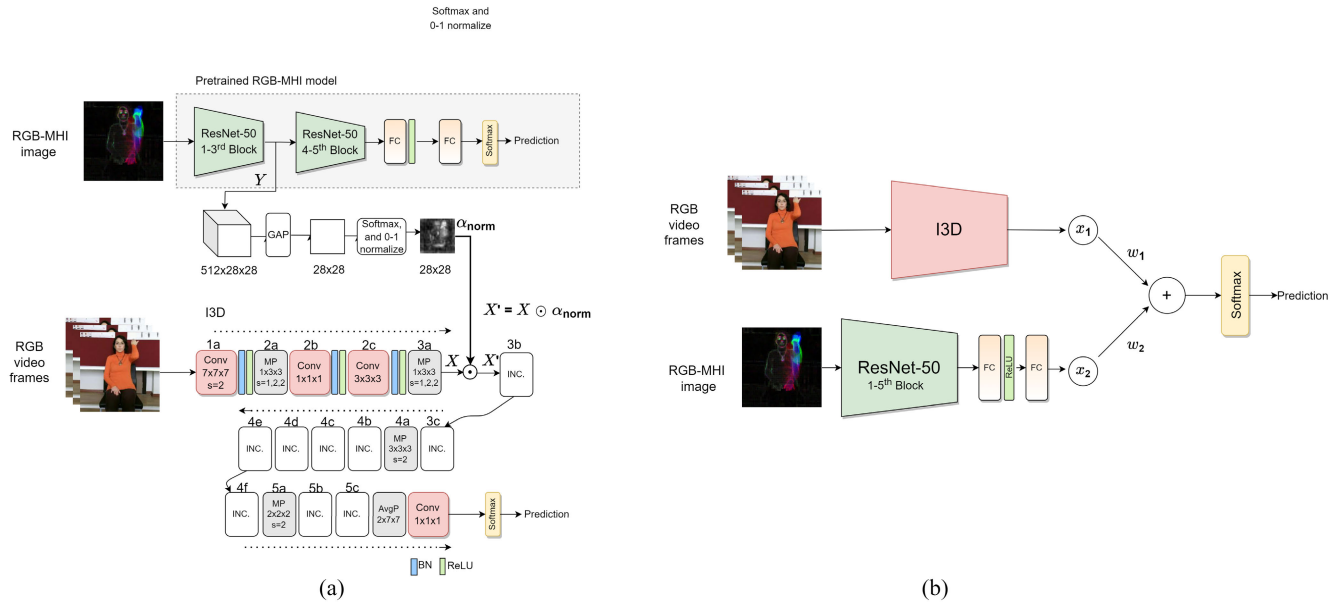

FIGURE 2. Proposed methods. (a) I3D with RGB-MHI attention (b) I3D+RGB-MHI fusion.

signer is always the closest person to the camera in this dataset and has the largest bounding box among people. We crop all video frames to be square by expanding the bounding box on both left and right sides (Fig. 3b, 3f). For 3D-CNN models, we fix the number of frames in all the videos to 32 frames. In isolated videos, the signer initially starts with a neutral position and returns to this position after performing the sign. For this reason, we skip some frames (maximum 10), which we determined according to the total number of frames in a video, from the beginning and the end of the stream. Then, we select 32 video frames from the remaining middle part by uniform sampling. On the other hand, while creating RGB-MHI images, we use all the video frames. After generating an RGB-MHI image, we crop it to be square using the bounding box information returned from Faster R-CNN of the corresponding video.

Moreover, in order to increase the variability, we apply data augmentation to RGB videos in the training and validation sets. In order to prevent the signers appear always in the center of the videos, we expand the bounding box from the left (Fig. 3c, 3g) or right (Fig. 3d, 3h), at first. In doing so, we also vertically shift the bounding box randomly (within a limited pixel) to allow the signer to be in different vertical positions. Finally, we also apply horizontal flip to these 3 versions of the data to accommodate our model for both hands. We resize all frames to 224×224.

B. TRAINING DETAILS

We implement all the models using PyTorch library [45]. In our models, some experiments are conducted in order to determine the optimum hyperparameters. The best results are obtained as follows. In the RGB-MHI model, modified C3D, and R(2+1)D models, we use Adam optimizer with an initial learning rate $1e-4$. In I3D-based models, we use SGD optimizer with an initial learning rate $1e-2$, and momentum 0.9. In S3D, we use Adam optimizer with an initial learning rate $1e-3$. We reduce the learning rate by 0.2 factor if no improvement is observed on the validation set for three epochs in all models. We repeat this process several times, and when there is no improvement anymore, we finally terminate the training.

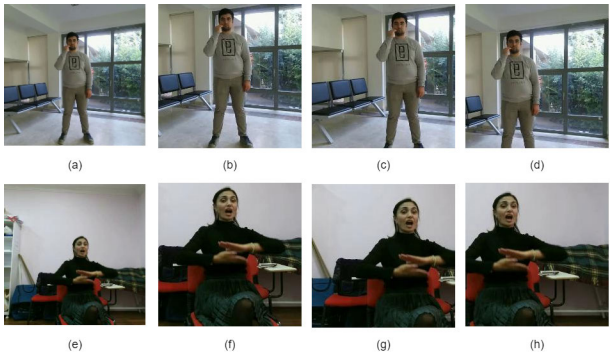

FIGURE 3. Examples of frame cropping and data augmentation of one standing, and one sitting signer from AUTSL dataset. (a, e) Original video frame. (b, f) After getting the bounding box with the Faster R-CNN [30], the bounding box is extended equally from the left and right to be a square. (c, g) The bounding box is expanded from the left side first. (d, h) The bounding box is expanded from the right side first. While expanding from the right and left, the bounding box is randomly shifted in the vertical direction.

C. EXPERIMENTS ON AUTSL

The results of our RGB-MHI model, which uses only one RGB-MHI image per video as an input, are reported in Table 1. In this approach, we modified the ResNet-50 model and made empirical observations on several versions with

TABLE 1. Ablation studies with RGB-MHI model on AUTSL dataset.

Method	Test Acc (%)
$Conv_1$ (64×112×112), MP (64×56×56), FC (1024)	38.52
$Conv_2$ (256×56×56), $Conv_{extra}$ (64×56×56), FC (1024)	45.90
$Conv_3$ (512×28×28), $Conv_{extra}$ (128×28×28), FC (1024)	60.04
$Conv_4$ (1024×14×14), FC (1024)	66.60
$Conv_4$ (1024×14×14), $Conv_{extra}$ (256×14×14), FC (1024)	68.77
$Conv_5$ (2048×7×7), FC (1024)	**71.95**
$Conv_5$ (2048×7×7), $Conv_{extra}$ (512×7×7), FC (1024)	69.85

The related feature sizes are shown in parentheses. $Conv_{1\text{-}5}$: Block names in ResNet-50, $Conv_{extra}$: Additional convolutional layer that we inserted, MP: Max Pooling, FC: Fully connected layer.

TABLE 2. Ablation studies with 3D-CNNs on AUTSL dataset.

Method	Params	Pretraining Dataset	Test Acc (%)
Modified C3D	27.7M	Sports1M [39]	83.04
I3D	12.5M	Kinetics400 [40]	87.72
I3D	12.5M	BSL-1K [41]	**89.30**
R(2+1)D-18	31.4M	Kinetics400 [40]	**89.67**
S3D	8.1M	Kinetics400 [40]	**89.73**

different depths. We achieved the best results when we used all the layers of ResNet-50. Compared to the baseline method, which is presented with the AUTSL dataset in [5], our best RGB-MHI model achieves 71.95% classification accuracy using a single image, which is higher than 2D-CNN and attention-based bidirectional LSTM model that takes entire video frames. Also, RGB-MHI model training is very fast since it only needs 1 motion history image, compared to 2D models where entire video frames are used as inputs. Moreover, since there are more than 200 different sign classes in the AUTSL dataset, 71.95% accuracy rate shows that a single RGB-MHI is capable of learning important discriminative features to represent isolated signs.

Since there is still a large room for performance improvement, we experiment with 3D-CNN based models using only RGB video frames. The results of the ablation studies with 3D-CNN architectures on AUTSL are provided in Table 2. Firstly, we experiment with the modified C3D model and achieve 83.04% accuracy. Then, we experiment with the I3D model that recently achieves successful performances on the recent large-scale sign language datasets, as well as in action recognition. In our first experiment, the I3D model is pre-trained on Kinetics-400 [40] action recognition dataset, and the second one is pre-trained on BSL-1K [41] British sign language dataset. We observe that when we transfer the features from BSL-1K, we obtain 1.58% higher accuracy. I3D that is pre-trained on BSL-1K, R(2+1)D-18, and S3D models all achieve above 89% recognition accuracy. We choose the I3D model as the baseline to use in our follow-up experiments since many recent state-of-the-art works in SLR also chose I3D as their baseline, e.g., WLASL [4] and MS-ASL [2].

TABLE 3. Accuracy results of the ablation studies on self-attention and RGB-MHI attention on AUTSL dataset.

Method	Test Acc (%)
I3D	89.30
Self-attention:	
I3D + 1 NL block after 3a	**89.94**
I3D + 2 NL blocks after 4a, 5a	89.38
I3D + 6 NL blocks after 4a, 4b, 4c, 4d, 4e, 4f	89.27
I3D + 9 NL blocks after 4a, 4b, 4c, 4d, 4e, 4f, 5a, 5b, 5c	89.57
RGB-MHI attention:	
I3D with RGB-MHI attention after 2a	89.38
I3D with RGB-MHI attention after 3a	**90.18**
I3D with RGB-MHI attention after 4a	89.14

In the follow-up works, we include the self-attention mechanisms to the selected 3D baseline model, i.e. I3D, to focus on related parts of the video frames. For this purpose, we insert different number of NL blocks [33] after different stages motivated from the original paper, specifically after 3rd, 4th, or 5th layers of I3D. In our experiments, we observe similar accuracy rates, but we get the best result (89.94%) when we insert 1 NL block after 3a block of I3D, only a slight improvement (0.64%) over the base model as seen in Table 3. Due to the memory constraints, we were able to add fewer NL blocks at lower level layers. However, we still obtain the best result when we add 1 NL block after the 3rd layer. We believe that the 3rd layer's larger spatial size is the main reason for that since it provides more spatial information.

1) I3D MODEL WITH RGB-MHI ATTENTION

We conduct some experiments with our proposed motion-based attention model. In this scope, we create attention weights from the feature maps taken from one of the middle-layer, i.e. 2nd or 3rd or 4th layer of RGB-MHI model. The sizes of the attention weights are 56 × 56, 28 × 28, 14 × 14, respectively. As in the experiments with the self-attention model, the best results in these experiments are obtained after the 3rd layer, that is, when the size of attention weights are 28 × 28. Our RGB-MHI based attention model performs 90.18% accuracy rate (Table 3), contributing 0.88% to the basic I3D model and 0.24% to the I3D model with self-attention.

2) I3D AND RGB-MHI FUSION MODEL

In our second approach, we fuse I3D and RGB-MHI models with a late fusion technique and obtain 91.13% with the weights of $w_1 = 0{,}6$; $w_2 = 0{,}4$. The empirical results show that RGB and RGB-MHI features are complementary to each other, thus the fusion of I3D and RGB-MHI model achieve better results than I3D with self-attention or I3D with RGB-MHI attention. Table 4 compares our proposed method variations.

We observe that our proposed I3D model and its RGB-MHI variants work successfully, invariant to the complex external motion arising from the background. In dynamic

TABLE 4. The results of our proposed method variations.

Method	Test Acc (%)
Without data augmentation:	
I3D	89.30
I3D with self attention	89.94
I3D with RGB-MHI attention	90.18
I3D+RGB-MHI fusion	**91.13**
I3D with RGB-MHI attention + fusion	**91.55**
With data augmentation (x6 more data):	
I3D	92.19
I3D fine tune with train+validation data	92.43
I3D+RGB-MHI fusion, $(w_1, w_2) = (0.7, 0.3)$	93.50
I3D+RGB-MHI fusion, $(w_1, w_2) = (0.6, 0.4)$	**93.53**
I3D+RGB-MHI fusion, $(w_1, w_2) = (0.5, 0.5)$	93.18

FIGURE 4. Sample videos that are correctly classified by the I3D, I3D with RGB-MHI attention, and I3D+RGB-MHI fusion models. Although the backgrounds contain moving people and swaying trees, all models could identify the correct sign without getting influenced by the cluttered background.

backgrounds, motion history images also summarize some undesirable (moving background) information (Fig. 4). However, our models can tolerate external motion in the background. In Fig. 4, we show examples that are correctly classified with all three models, i.e. I3D, I3D with RGB-MHI attention, and I3D+RGB-MHI fusion. I3D model alone already performs well in these samples, hence we did not observe additional help from RGB-MHI images in correct classifications for these samples. Still, I3D with RGB-MHI model variants also correctly classify these challenging samples, even RGB-MHI images contain background motion.

On the other hand, the overall test scores depict that the RGB-MHI model helps I3D to better classify the challenging signs in the AUTSL dataset. Note the accuracy rates of I3D, I3D with RGB-MHI attention, and I3D+RGB-MHI fusion in Table 4, which are 89.30%, 90.18%, 91.13%, respectively. When we analyze test results, we observe that while some videos are misclassified with I3D, they are classified correctly when the RGB-MHI features are incorporated. We observe that using RGB-MHI features helps to correctly classify some challenging signs where the hand shape is quite similar, yet there are differences in hand

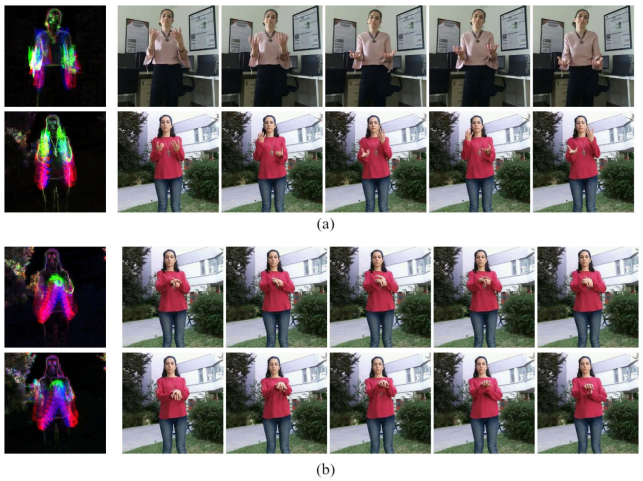

FIGURE 5. Sample videos that are misclassified by the I3D model and correctly classified by the I3D with RGB-MHI attention and I3D+RGB-MHI models. Each row contains some sample video frames and the corresponding RGB-MHI image. (a-first row) "agir: heavy" is misclassified with the I3D model as "ocak: oven" (second row). As seen in the video frames, hand shapes are very similar, but there is a difference in the movement of hands. In the "heavy" sign, both hands are lowered from above, while in the "oven", the hands move in opposite directions. (b-first row) "yavas: slow" sign is misclassified as "eldiven, gloves" (second row). These two signs are very similar; there is only a slight difference in hand orientation and movement.

movements (Fig. 5). Note that in these test samples, the background environments are not simple.

3) I3D WITH RGB-MHI ATTENTION AND FUSION MODEL

Finally, we investigated the contribution of both the RGB-MHI attention and RGB-MHI model fusion. When we fuse the I3D with RGB-MHI attention model and the RGB-MHI model, we obtain 91.55%. Considering the trade-off between the model complexity and accuracy rates, we selected the I3D+RGB-MHI fusion model as our final model.

In order to increase the generalization capability of the model, we train our base I3D model with the augmented data that is described in Section IV-A. In this experiment, the classification accuracy has been improved from 89.30% to 92.19% when 6 times more train data, which is generated with data augmentation, has been used. In an SLR challenge [24] on the AUTSL dataset, participants are allowed to use validation labels to fine-tune their models. Therefore, we fine-tune our model with the union of training and validation sets and obtain 92.43% classification accuracy. In this experiment, the initial learning rate is set to $2e-3$, instead of $1e-2$. Since there is no validation set in this experiment, we stop fine-tuning when the training loss is reduced to the same level as in the previous run. Finally, we ensemble it with the RGB-MHI model using a predefined set of fusion weights. Since I3D achieves a higher accuracy rate than the RGB-MHI model, we give higher or equal weights to I3D model and we experiment with three different weight pairs for fusion: (0.7, 0.3), (0.6, 0.4), (0.5, 0.5), respectively. The

TABLE 5. Comparison of the literature on AUTSL.

Method	Modalities	Test Acc (%)
SAM-SLR [23]	RGB, skeleton joints, bones, joint motion, bone motion, skeleton features, optical flow	98.42
MS-G3D [46]	RGB, skeleton joints, bones, joint motion, bone motion	96.15
I3D-VLE-transformer [26]	RGB, hand crops, pose, location vectors	95.46
I3D+RGB-MHI fusion	**RGB, RGB-MHI image**	**93.53**
VTN-PF [27]	RGB, hand crops, location vectors, skeleton	92.92
OpenPose+Holistic [47]	Skeleton	81.93
2DCNN+Attentioned BLSTM [5]	RGB	49.22

TABLE 6. Performances of the proposed methods on BosphorusSign22k.

Model	Pretrained Dataset	Test Acc (%)
RGB-MHI	ImageNet	75.77
I3D	BSL-1K	91.64
I3D	AUTSL	92.66
I3D fine tune with train+validation data	AUTSL	94.47
I3D+RGB-MHI fusion	AUTSL, ImageNet	**94.83**

TABLE 7. Comparison of the literature on BosphorusSign22k.

Method	Modalities	Test Acc (%)
MC3-18 Spatio-Temporal Sampling [28]	RGB, cropped hands, cropped face	94.94
I3D+RGB-MHI	**RGB, RGB-MHI image**	**94.83**
IDT [6]	HOG, HOF, MBH	88.53
MC3-18 Spatio-Temporal Sampling [28]	RGB	86.91
MC3-18 [6]	RGB	78.85

fusion model achieves approximately 1% better, and our best I3D+RGB-MHI model performs with 93.53% accuracy, using (0.6, 0.4) weight pairs.

Table 5 contains the state-of-the-art model performances on AUTSL in the literature. All recent models perform considerably higher than the baseline 2D-CNN+LSTM based model. Most of the works use multiple modalities, such as skeleton joints, optical flow, hand crops, etc., and model ensembles to increase the classification accuracies. Our model performs comparable to the state-of-the-art models, although we only used RGB frames and the RGB-MHI image that we generate from RGB data.

D. EXPERIMENTS ON BosphorusSign22k

We evaluate our I3D+RGB-MHI fusion model on another large-scale sign language dataset, BosphorusSign22k [6], since our I3D+RGB-MHI fusion model achieves a better result than I3D with our RGB-MHI attention model. RGB-MHI model alone classifies the signs in this dataset with 75.77% accuracy. Following that, we conduct similar experiments with the I3D model using augmented train data. In our first experiment, we use the BSL-1K pre-trained version of the I3D model and obtained 91.64% accuracy. Next, in order to analyze the results of the transfer learning from different datasets, we use AUTSL-trained parameters instead of BSL-1K in the same model. In this case, we obtain slightly better accuracy, i.e. 92.66%. We first thought that pretraining with AUTSL is more helpful since both datasets are TSL datasets, and hence contain similar signs (although performed in different contexts and with different signers). However, we realized from the works of [46] that pretraining with AUTSL is effective in Spanish Sign Language as well. In Vazquez-Enriquez *et al.*'s work [46], the authors show the effect of transferring data from different sign language datasets. In their experiments, they worked on LSE-Lex 40 [48] Spanish Sign Language dataset; they trained their model from scratch or applied transfer learning to models that were trained with AUTSL or WLASL datasets. They observed that the best results are obtained with an AUTSL pre-trained model. In experiments conducted on both BosphorusSign22k and LSE-Lex 40 datasets, higher accuracy rates are obtained when initial parameters are transferred from the AUTSL dataset, indicating that AUTSL provides a good initialisation for other sign language datasets. Therefore, we choose the AUTSL pre-trained I3D model as our baseline in these experiments. Then, we fine-tune the I3D model with the union of the train and validation datasets. Since there are only 4 signers in the training data, an additional 1 person in the validation set contributed to the increase in generalization of the model, and we obtained 94.47% accuracy. Finally, we fuse our I3D and RGB-MHI model with the coefficients of $w_1 = 0.6$; $w_2 = 0.4$, and obtained 94.83% accuracy. Our results are provided in Table 6.

Table 7 compares model performances on BosphorusSign22k in the literature. As seen in the table, our proposed I3D+RGB-MHI fusion model achieves competitive performance, 94.83%, with the state-of-the-art; following the best performing method very closely. Notice that in our proposed model, no explicit part segmentation is included as a separate modality. The context provided with the RGB-MHI image helps the 3D model to focus on relevant parts of the sign action.

V. CONCLUSION

In the isolated SLR literature, the methods that train different models with multiple modalities and use ensembles of these models with various fusion techniques achieve more successful results. Although performing significantly well, such models are complex and usually demand more computational resources. In this research, our aim is to utilize only the RGB data modality to avoid additional modalities without sacrificing the model performances significantly. Two methods are proposed in this regard; in the first

approach, rather than explicitly segmenting the hands or face, a model that focuses on motion patterns is proposed with the RGB-MHI images. For the second approach, RGB and RGB-MHI features are fused in a late fusion technique, before the prediction. Compared to the literature, our proposed models perform comparably to the state-of-the-art models that use multiple data modalities; and provide strong baselines for the models that use RGB data on two large-scale TSL benchmarks. As future work, we plan to investigate the effectiveness of the proposed models in the continuous SLR domain.

ACKNOWLEDGMENT

The numerical calculations reported in this paper were partially performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

REFERENCES

[1] *World Federation of the Deaf*. Accessed: Jul. 16, 2021. [Online]. Available: https://wfdeaf.org/our-work/

[2] H. R. V. Joze and O. Koller, "MS-ASL: A large-scale data set and benchmark for understanding American sign language," 2018, *arXiv:1812.01053*.

[3] J. Huang, W. Zhou, H. Li, and W. Li, "Attention-based 3D-CNNs for large-vocabulary sign language recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2822–2832, Sep. 2019.

[4] D. Li, C. R. Opazo, X. Yu, and H. Li, "Word-level deep sign language recognition from video: A new large-scale dataset and methods comparison," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 1459–1469.

[5] O. M. Sincan and H. Y. Keles, "AUTSL: A large scale multi-modal Turkish sign language dataset and baseline methods," *IEEE Access*, vol. 8, pp. 181340–181355, 2020.

[6] O. Özdemir, A. A. Kındıroğlu, N. C. Camgöz, and L. Akarun, "Bosphorus-sign22k sign language recognition dataset," in *Proc. LREC 9th Workshop Represent. Process. Sign Languages, Sign Lang. Resour. Service Lang. Community, Technol. Challenges Appl. Perspect.*, 2020, pp. 181–188.

[7] H. Hu, W. Zhou, J. Pu, and H. Li, "Global-local enhancement network for NMFs-aware sign language recognition," 2020, *arXiv:2008.10428*.

[8] O. Koller, "Quantitative survey of the state of the art in sign language recognition," 2020, *arXiv:2008.09918*.

[9] A. Z. Shukor, M. F. Miskon, M. H. Jamaluddin, F. B. Ali@Ibrahim, M. F. Asyraf, and M. B. B. Bahar, "A new data glove approach for Malaysian sign language detection," *Proc. Comput. Sci.*, vol. 76, pp. 60–67, Jan. 2015.

[10] Z. Katılmış and C. Karakuzu, "ELM based two-handed dynamic Turkish sign language (TSL) word recognition," *Expert Syst. Appl.*, vol. 182, Nov. 2021, Art. no. 115213.

[11] M. Al-Qurishi, T. Khalid, and R. Souissi, "Deep learning for sign language recognition: Current techniques, benchmarks, and open issues," *IEEE Access*, vol. 9, pp. 126917–126951, 2021.

[12] P. Jangyodsuk, C. Conly, and V. Athitsos, "Sign language recognition using dynamic time warping and hand shape distance based on histogram of oriented gradient features," in *Proc. 7th Int. Conf. Pervasive Technol. Rel. Assistive Environ.*, May 2014, pp. 1–6.

[13] A. Memis and S. Albayrak, "Turkish sign language recognition using spatio-temporal features on Kinect RGB video sequences and depth maps," in *Proc. 21st Signal Process. Commun. Appl. Conf. (SIU)*, Apr. 2013, pp. 1–4.

[14] G. A. Rao and P. Kishore, "Selfie sign language recognition with multiple features on adaboost multilabel multiclass classifier," *J. Eng. Sci. Technol.*, vol. 13, no. 8, pp. 2352–2368, 2018.

[15] L. Pigou, A. Van Den Oord, S. Dieleman, M. Van Herreweghe, and J. Dambre, "Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video," *Int. J. Comput. Vis.*, vol. 126, nos. 2–4, pp. 430–439, 2015.

[16] O. Mercanoglu Sincan, A. O. Tur, and H. Yalim Keles, "Isolated sign language recognition with multi-scale features using LSTM," in *Proc. 27th Signal Process. Commun. Appl. Conf. (SIU)*, Apr. 2019, pp. 1–4.

[17] H. Luqman and E.-S.-M. El-Alfy, "Towards hybrid multimodal manual and non-manual Arabic sign language recognition: MArSL database and pilot study," *Electronics*, vol. 10, no. 14, p. 1739, Jul. 2021.

[18] R. Rastgoo, K. Kiani, and S. Escalera, "Video-based isolated hand sign language recognition using a deep cascaded model," *Multimedia Tools Appl.*, vol. 79, nos. 31–32, pp. 22965–22987, Aug. 2020.

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

[20] C. C. dos Santos, J. L. A. Samatelo, and R. F. Vassallo, "Dynamic gesture recognition by using CNNs and star RGB: A temporal information condensation," *Neurocomputing*, vol. 400, pp. 238–254, Aug. 2020.

[21] J. Imran and B. Raman, "Deep motion templates and extreme learning machine for sign language recognition," *Vis. Comput.*, vol. 36, no. 6, pp. 1233–1246, Jun. 2020.

[22] P. Barros, G. I. Parisi, D. Jirak, and S. Wermter, "Real-time gesture recognition using a humanoid robot with a deep neural architecture," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, Nov. 2014, pp. 646–651.

[23] S. Jiang, B. Sun, L. Wang, Y. Bai, K. Li, and Y. Fu, "Skeleton aware multi-modal sign language recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3413–3423.

[24] O. M. Sincan, J. C. S. Jacques, S. Escalera, and H. Y. Keles, "ChaLearn LAP large scale signer independent isolated sign language recognition challenge: Design, results and future research," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3472–3481.

[25] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.

[26] I. Gruber, Z. Krnoul, M. Hruz, J. Kanis, and M. Bohacek, "Mutual support of data modalities in the task of sign language recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3424–3433.

[27] M. De Coster, M. Van Herreweghe, and J. Dambre, "Isolated sign recognition from RGB video using pose flow and self-attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3441–3450.

[28] C. C. Gökçe, O. Özdemir, A. A. Kındıroğlu, and L. Akarun, "Score-level multi cue fusion for sign language recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 294–309.

[29] R. Rastgoo, K. Kiani, and S. Escalera, "Multi-modal deep hand sign language recognition in still images using restricted Boltzmann machine," *Entropy*, vol. 20, no. 11, p. 809, 2018.

[30] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[31] B. Shi, A. M. D. Rio, J. Keane, D. Brentari, G. Shakhnarovich, and K. Livescu, "Fingerspelling recognition in the wild with iterative visual attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5400–5409.

[32] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5533–5541.

[33] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[34] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[35] E. C. Alp and H. Y. Keles, "Action recognition using MHI based Hu moments with HMMs," in *Proc. IEEE EUROCON 17th Int. Conf. Smart Technol.*, Jul. 2017, pp. 212–216.

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

[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[38] 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. (ICCV)*, Dec. 2015, pp. 4489–4497.
[39] 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 Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1725–1732.
[40] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*.
[41] S. Albanie, G. Varol, L. Momeni, T. Afouras, J. S. Chung, N. Fox, and A. Zisserman, "BSL-1K: Scaling up co-articulated sign language recognition using mouthing cues," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2020, pp. 35–53.
[42] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 305–321.
[43] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.
[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[45] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, and A. Desmaison, "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32. Vancouver, BS, Canada, Dec. 2019, pp. 8026–8037.
[46] M. Vazquez-Enriquez, J. L. Alba-Castro, L. Docio-Fernandez, and E. Rodriguez-Banga, "Isolated sign language recognition with multi-scale spatial-temporal graph convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3462–3471.
[47] A. Moryossef, I. Tsochantaridis, J. Dinn, N. C. Camgoz, R. Bowden, T. Jiang, A. Rios, M. M´uller, and S. Ebling, "Evaluating the immediate applicability of pose estimation for sign language recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 3434–3440.
[48] L. Docío-Fernández, J. L. Alba-Castro, S. Torres-Guijarro, E. Rodríguez-Banga, M. Rey-Area, A. Pérez-Pérez, S. Rico-Alonso, and C. G. Mateo, "Lse_uvigo: A multi-source database for Spanish sign language recognition," in *Proc. LREC 9th Workshop Represent. Process. Sign Lang., Sign Lang. Resour. Service Lang. Community, Technol. Challenges Appl. Perspect.*, 2020, pp. 45–52.

OZGE MERCANOGLU SINCAN (Member, IEEE) was born in Ankara, Turkey, in 1989. She received the B.S., M.S., and Ph.D. degrees in computer engineering from Ankara University, Turkey, in 2012, 2016, and 2021, respectively.

Since 2013, she has been a Research Assistant at the Department of Computer Engineering, Ankara University. She is a member of the Computer Vision and Machine Learning (CVML) Laboratory, Department of Computer Engineering. She is one of the co-organizers of a Sign Language Recognition Workshop and Challenge at CVPR 2021. She has been working in some research projects related to deep learning, in collaboration with the industry. Her research interests include computer vision and deep learning, particularly in the sign language recognition problems.

HACER YALIM KELES was born in Ankara, Turkey, in 1979. She received the B.Sc., M.Sc., and Ph.D. degrees in computer engineering from Middle East Technical University, Turkey, in 2002, 2005, and 2010, respectively.

From 2000 to 2007, she worked as a Researcher at the Scientific and Technological Research Council of Turkey (TUBITAK). From 2013 to 2020, she worked as an Assistant Professor at the Department of Computer Engineering, Ankara University, where she was an Associate Professor, from 2020 to 2021. She is currently an Associate Professor at the Department of Computer Engineering, Hacettepe University, Turkey. She is one of the co-organizers of a recently held workshop and challenge on Sign Language Recognition under CVPR 2021 Conference (SLRITW). Her research interests include computer vision and machine learning, particularly in sign language recognition and generative adversarial networks. She is also interested in pattern recognition problems from various domains, including medical science and remote sensing areas.

Dr. Keles's Ph.D. thesis received the Thesis of the Year Award by Middle East Technical University Prof. Dr. Mustafa Parlar Education and Research Foundation, in 2010. In 2010, she received a Research and Development Grant from the Ministry of Industry and Trade of Turkey and established her own research and development company. Her follow-up project SOYA was funded by TUBITAK, in 2011, and later awarded by TUBITAK as one of the ten best venture projects in the country. She is the first woman who took this grant in Turkey.