IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received August 11, 2020, accepted September 5, 2020, date of publication September 11, 2020,
date of current version September 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3023599

Infrared and 3D Skeleton Feature Fusion for RGB-D Action Recognition

ALBAN MAIN DE BOISSIERE AND RITA NOUMEIR, (Member, IEEE)
Laboratoire de Traitement de l'Information en Santé, École de Technologie Supérieure, Montreal, QC H3C 1K3, Canada

Corresponding author: Alban Main de Boissiere (alban.main-de-boissiere.1@ens.etsmtl.ca)

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada, Prompt Québec, and in part by Aerosystems International Inc.

ABSTRACT For skeleton-based action recognition from depth cameras, distinguishing object-related actions with similar motions is a difficult task. The other available video streams (RGB, infrared, depth) may provide additional clues, given an appropriate feature fusion strategy. We propose a modular network combining skeleton and infrared data. A pre-trained 2D convolutional neural network (CNN) is used as a pose module to extract features from skeleton data. A pre-trained 3D CNN is used as an infrared module to extract visual features from videos. Both feature vectors are then fused and exploited jointly using a multilayer perceptron (MLP). The 2D skeleton coordinates are used to crop a region of interest around the subjects for the infrared videos. Infrared is favored over RGB, as it is less affected by illumination conditions and usable in the dark. We are the first to combine infrared and skeleton data. We evaluate our method on the NTU RGB+D dataset, the largest dataset for human action recognition from depth cameras. We perform extensive ablation studies. In particular, we show the strong contributions of our cropping strategy and pre-training on action classification accuracy. We also test various feature fusion schemes. Feature sum on an element-wise level yields the best results. Our method achieves state-of-the-art performances on NTU RBG+D.

INDEX TERMS Action recognition, depth cameras, feature extraction, gesture recognition, infrared, skeleton, video understanding.

I. INTRODUCTION

Human action recognition is the task of recognizing an activity performed by one or more subjects inside a segmented sequence. Recent years have witnessed successful deep architectures [5], [9], [23], [39], [52], [62] with promising results on benchmark datasets [3], [38].

Consumer-grade depth cameras such as Intel RealSense [22] and Microsoft Kinect [66] coupled with advanced human pose estimation algorithms [43] have allowed 3D skeleton data to be obtained in real-time. Key joints of the human body are extracted to a 3D space, providing a high-level representation of an action. Skeleton data are robust to surrounding environment, illumination variations and may be generalized to various viewpoints [1], [11], [28], [29], [35], [57]. Earlier works have indicated that key joints are powerful descriptors for human motion [18]. The low dimensionality and high representation power make skeleton data a prime input for action recognition tasks.

Opening the door for new action recognition algorithms, those are broadly categorized into RGB and 3D skeleton approaches. However, it has been demonstrated that visual and skeleton inputs can work in symbiosis [36]. Actions with similar body motion, such as writing versus typing on a keyboard, prove difficult to classify with skeleton data only. In this respect, skeleton data might benefit from the visual clues of RGB streams.

Depth cameras offer four different data streams: RGB, depth, infrared (IR) videos, and 3D skeleton. To our knowledge, infrared videos from depth cameras have never been used as an input source for action recognition. We argue that the lack of large scale datasets proposing IR videos in addition to the other streams is in part responsible. Moreover, RGB and IR images are quite similar, the former offering a richer representation of a scene, therefore, making it a better candidate. However, IR is usable in the dark, which is viable for security applications when skeleton data are

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

insufficient. The recent introduction of large scale datasets like NTU RGB+D [38] and PKU-MMD [31] containing IR videos motivates the evaluation of methods using this stream. Video understanding is a well-studied computer vision task. But modeling spatiotemporal features and long-term dependencies remains an issue.

Another challenge in video action classification is the volume of information. To reduce the complexity of the videos, downscaling the frames is often employed but also comes with a decrease in the quality of the information. Moreover, discriminating clues may only occur in a small portion of the frames, becoming undetectable in the process [51]. An alternative proposal is to focus on regions of interest. Visual attention models are capable of focusing on important cues and disregard other areas [4], [34], [40].

In this work, we intend to address the difficulty of differentiating actions with similar motions with an additional visual stream insensible to illumination conditions. Furthermore, we evaluate the potential of IR videos as a standalone source. We propose a model fusing video and pose data (FUSION). Pose has a double purpose. It is used as an input stream in its own right and also conditions the IR sequences, providing a crop around the subjects, facilitating the classification. The general outline of the network is illustrated Fig. 1.

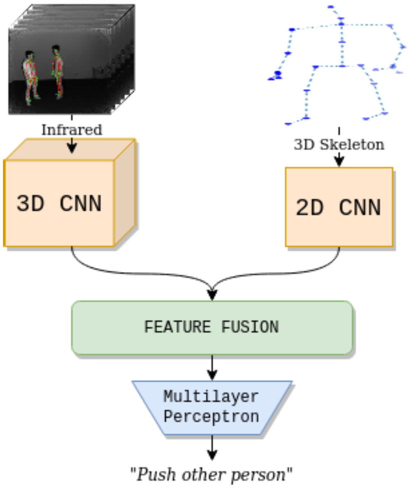

FIGURE 1. Our model uses a 2D CNN for pose data and a 3D CNN for IR sequences. Features from both modules are then fused and studied jointly via an MLP. Training is done in end-to-end fashion.

The pose network is an 18-layer ResNet [12] taking as input the entire skeleton sequence. The sequence is mapped to an RGB image which is then rescaled to fit the input size of the CNN. The IR network is a ResNet (2+1)D (R(2+1)D) [50] where a fixed number of random frames taken from evenly spaced subsequences are used as inputs. The features of each module are then fused before proposing a final classification with a multilayer perceptron (MLP).

Our main contributions are as follows:

- We demonstrate the importance of IR streams from depth cameras for human action recognition.
- We propose a fusion network taking skeleton and IR sequences as inputs. Utilizing those two steams conjointly has never been attempted before.
- We perform extensive ablation studies. We isolate different modules of our model and study their representation power. We also evaluate the importance of data augmentation, transfer learning, 2D-skeleton conditioned IR sequences, IR sequence length, and various feature fusion strategies on the accuracy score.
- We achieve state-of-the-art results compared to methods using different streams.

Codes, documentation, and supplementary materials can be found on the project page.[1]

II. RELATED WORK

A. SKELETON-BASED APPROACHES

Human action recognition has received a lot of attention due to its high-level representation and powerful discriminating nature. Traditional approaches focus on handcrafted features [16], [53], [56]. These could be the dynamics of joint motion, the covariance matrix of joint trajectories [16] or the representation of joints in a Lie group [53]. Design choices prove challenging and result in suboptimal results. Recent deep-learning methods report improved accuracy. There exist three main frameworks: sequence-based models, image-based models, and graph-based models.

Sequence models exploit skeleton data as time series of key joints which are then fed to recurrent neural networks (RNN) [9], [26], [32], [38], [47], [55], [64]. Part-aware long short-term memory (LSTM) RNN [38] uses different memory cells for different regions of the body, then fuses them for the final classification. Similarly in [9], a bidirectional RNN studies separate body parts individually in earlier levels and jointly deeper on. In an effort to model simultaneously time and spatial dependencies, Liu *et al.* propose a 2D recurrent model [32]. Recurrent models are now part of the early deep learning efforts for skeleton-based action recognition. Vastly improving upon the results of the traditional methods, they remain insufficient. The sequence length has to be fixed during training which is not ideal and requires a sampling strategy. Moreover, sequence models tend to be much slower than their image-based counterparts.

Image models represent skeleton data as 2D images which are then used as inputs for convolutional neural networks (CNN) [8], [21], [23], [27]–[29], [33], [59]. An intuitive method is to assign the x, y and z coordinates of a skeleton sequence to the channels of an RGB image [8], [27]. Each joint corresponds to a row and each frame to a column, or inversely. Pixel intensity is then normalized between 0 and 255 based on maximal coordinates value of the dataset [8] or sequence [27]. Other works utilize the relative coordinates between joints to generate multiple images [21]. Similarly, some works project the 3D coordinates on orthogonal 2D planes [28], [29] and encode the trajectories into a hue,

[1] https://github.com/adeboissiere/FUSION-human-action-recognition

saturation, value (HSV) space [59]. A pre-trained model over ImageNet [6] is leveraged. A similar approach is used in [13]. More recent works focus on view-invariant transformations [20], [33] or networks [65] with improved results. In [23], a temporal convolutional network is deployed with interpretability of the results as a major objective. CNNs are able to learn from entire sequences rather than sampled frames. The image generated from the skeleton sequence is resized to accommodate the fixed input shape of the CNN. This means an entire sequence can be used at once, which is an advantage compared to recurrent methods.

Graph neural networks have received a lot of attention as of late due to their effective representation of skeleton data [61]. There exist two main graph model architectures: graph neural networks (GNN), and graph convolutional networks (GCN), which aim to generalize traditional convolutional networks. Spatial GCNs leverage the convolution operator for each node using its nearest neighbors [45]. Yan *et al.* [62] make the best of the graph representation to learn both spatial and temporal features. Li *et al.* generalize the graph representation to actional and structural links [30]. In [44], a temporal attention mechanism is adopted to enhance the classification while exploring the co-occurrence relationship between spatial and temporal domains. In [42], the length and direction of bones are used in addition to joint coordinates while adapting the topology of the graph. Shi *et al.* represent skeleton data as a directed acyclic graph based on kinematic dependencies of joints and bones [41]. GCNs report the current state-of-the-art results on benchmark datasets. However, carefully designed CNNs show comparable results [65]. Also, CNNs can be pre-trained on other large scale datasets which improves the performances of image-based skeleton action recognition models [65]. To our knowledge, an ImageNet [6] style transfer learning is impractical for GCNs.

B. RGB-BASED VIDEO CLASSIFICATION

Traditional approaches focus on handcrafted features in the form of spatiotemporal interest points. Among those, improved Dense Trajectories (iDT) [54], which uses estimated camera movements for feature correction, is considered the state of the art. After the widespread use of deep learning on single images, many attempts have been made to propose benchmarks for video classification.

Soon after [54], two breakthrough papers [19], [46] would form the backbone of future efforts. In [19], Karpathy *et al.* explore different ways of fusing temporal information using pre-trained 2D CNNs. In [46], handcrafted features, in the form of optical flow, are used symbiotically with the raw video. Two parallel networks compute spatial and temporal features. A few drawbacks include the inability to effectively capture long-range temporal information and the heavy calculations required to compute optical flow.

Later research propositions fall into five frameworks:

- 2D CNN followed by RNN network [7]
- 3D CNN [5], [49], [63]
- Two-Stream 2D CNN [10]
- 3D-Fused Two-Stream [10]
- Two-Stream 3D CNN [3], [52]

Heavy networks and computations of handcrafted features, as well as the absence of a benchmark for long-term temporal features, remain an issue. In [50], Tran *et al.* explore different forms of spatiotemporal convolutions and their impact on video understanding. A (2+1)D convolution block separating spatial and temporal filters allows for a greater nonlinearity compared to a standard 3D block with an equivalent number of parameters, as illustrated Fig. 2. Separating convolutions yields state-of-the-art results on benchmark datasets such as Sports-1M [19], Kinetics [3], UCF101 [48] and HMDB51 [25].

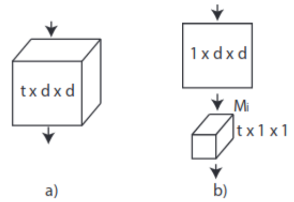

FIGURE 2. a) A standard 3D convolution operator. b) A factorized (2+1)D convolution operation with an additional nonlinear activation function in between. Illustration courtesy of [50].

C. MIXED INPUTS ACTION RECOGNITION

Depth cameras provide different streams, or in other words, different representations of the same action. Some works have attempted to improve classification by combining streams. It can be argued that skeleton-based approaches prove most effective at discriminating actions with broad movements. However, for actions involving similar joint positions and trajectories, such as reading vs. playing on a phone, skeleton-based models do not perform as well. Visual streams can provide important cues such as the type of object held. RGB and depth streams have been studied extensively. However, to our knowledge, we are the first to use IR data from depth cameras for action recognition.

In [15], [39], [58] the complementary role of RGB and depth is demonstrated. In [67], pose, motion, and raw RGB images are inputted in 3 parallel 3D CNNs. Although visual information greatly improves upon the pose baseline, results are comparable with the then state-of-the-art methods using only skeleton data. Pose data can be utilized to extract regions of interest around joints or body parts [2], [14], [36]. In [36], human-object interactions are modeled using both skeleton and depth data. An end-to-end network is proposed to learn view-invariant representations of skeleton data and held objects. Once again, visual information increases the accuracy but the results do not justify the complexity of a fusion approach compared to the other skeleton-only approaches of the time. The same year, Baradel *et al.* use RGB and skeleton data jointly in a pertinent way [2]. Pose information is used as an input but also conditions the RGB stream. The 3D skeleton data are projected onto the RGB sequences to effectively extract crops around the hands of the subject, serving as another input. The RGB stream thus provides important clues about an object held and inter-subject interactions, significantly improving the results. This work shows that not all body parts need to be focused on, unlike the approach in [36]. But this requires as many streams as there are hands, which is memory inefficient. Furthermore, when the hands are close together, the information provided may be redundant. Alternatively in [52], a region of interest is created using motion fields from RGB videos. An additional region from the body is extracted using motion saliency. The advantages of this method are that depth data are not required and the attention mechanism role of the saliency map. But for almost motionless actions, the region extraction should not perform as well.

We propose a similar approach to [2], [36] and [52] in which the 3D skeleton data provide a crop around the subjects, alleviating the need for a spatial attention mechanism. A single crop is necessary, even when multiple subjects are interacting, which relaxes the memory needs. Comparably, a faster R-CNN [37] can be used to crop a region of interest around the subjects on depth images [60], but requires a powerful graphics processing unit to be usable in real-time.

III. PROPOSED MODEL

We design a deep neural network using skeleton and IR data, called "Full Use of Infrared and Skeleton in Optimized Network" (FUSION). The network consists of two parallel modules and an MLP. One module interprets skeleton data, the other IR videos. The features extracted from each stream are then fused using different strategies (average, sum, multiplication, max, convolution, concatenation). The MLP is used as the final module and outputs a probability density. The network is trained in end-to-end fashion by optimizing the classification score.

We note a skeleton sequence $\mathbf{S} = \{\mathbf{S}_{j,t,k}\}$ where j denotes a joint index, t a frame index and k a coordinate axis (X, Y and Z). We note $\mathbf{I} = \{\mathbf{I}_t\}$ a sampled IR sequence, as detailed section III-B3, where t is taken between $\{1, .., T\}$, with T the number of sampled frames.

In the following sections, we present the individual modules of our FUSION model: a 2D CNN as the pose module, a 3D CNN as the IR module, and an MLP as the stream fusion module.

A. POSE MODULE

A skeleton sequence requires careful treatment for optimal results. First, a skeleton sequence is normalized to be position invariant, meaning the distance between the subject and the camera is accounted for. The sequence is then transcribed to an RGB image, with multiple subjects interactions in mind. The handcrafted RGB image is then fed to a 2D CNN.

1) PRIOR NORMALIZATION STEP

Each skeleton sequence is normalized by translating the global coordinate system of the camera to a local coordinate system corresponding to a key joint of the main subject. We choose the middle of the spine as the new origin. This is illustrated Fig. 3.

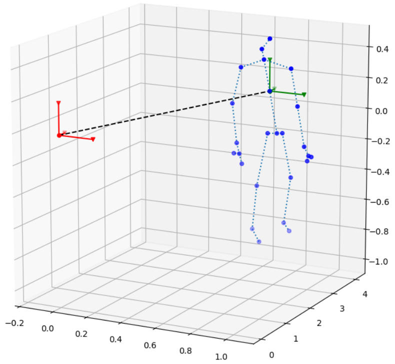

FIGURE 3. In red the coordinate system of the camera, in green the new coordinate system corresponding to the middle of the spine of the main subject for the first frame of the sequence, in blue the skeleton of the main subject, in black the translation vector.

We adopt a sequence-wise normalization. In other words, the translation vector is computed for the first frame and applied to each subsequent frame, meaning the subject may move away from the new local coordinate system, as follows:

$$\mathbf{S}' = \mathbf{S}_{:,:,:} - \mathbf{S}_{1,0,:}. \tag{1}$$

where $\mathbf{S}'$ is the normalized skeleton sequence, $j = 1$ corresponds to the middle of the spine for the Kinect 2 skeleton [66]. The ":" notation signifies that all values are considered across this dimension.

2) SKELETON DATA TO SKELETON 2D MAPS

A skeleton sequence is mapped to an image similar to [8], a skeleton map. Each coordinate axis, X, Y and Z, is attributed to each channel of an RGB image. Each key joint corresponds to a row while the columns represent the different frames.

We apply a dataset-wise normalization [8]. We note c_{min} and c_{max} the minimal and maximal values of the coordinates after the normalization step for the entire training set. As such, c_{min} and c_{max} are not influenced by the validation and testing sets. The pixels of the skeleton map are recalculated using a min-max strategy in the [0, 1] range, as follows:

$$\mathbf{M} = \frac{\mathbf{S}' - c_{min}}{c_{max} - c_{min}}. \tag{2}$$

where $\mathbf{M} = \{\mathbf{M}_{j,t,k}\}$ is the normalized skeleton map with k both the coordinate axis and the image channel.

To accommodate for the fixed input size of the 2D CNN, the skeleton map is resized to a standard size.

3) MULTI-SUBJECT STRATEGY

Our network is scalable to multiple subjects. We concatenate the different skeleton maps across the joint dimension. With

J being the total number of joints, the first J rows correspond to the first subject, the subsequent J rows to subject 2, etc. We limit the number of subjects to two, corresponding to the maximum of the NTU RGB+D dataset [38]. Nonetheless, this method may be generalized to a greater number of subjects. Should the skeleton sequence comprise only one subject, the J rows of the second subject are set to zero.

In the case of multiple subjects, the coordinates of the latter are translated to the local coordinate system of the main subject (Fig. 4).

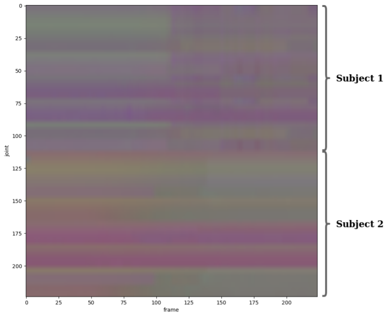

FIGURE 4. Skeleton map of two subjects. The joints of the two subjects are concatenated across a dimension, then stacked over time. The created image is reshaped to the fixed CNN input size.

The advantages of our method are manifold. Firstly, this alleviates the need for individual networks for different subjects. Secondly, this representation allows for a second subject to still intervene if its skeleton is detected after the first frame. Thirdly, the distance information is kept as each subject coordinates are translated to the local coordinate system of the first subject. Lastly, the skeleton map is resized to a standard size to accommodate for the fixed input size of the pose module. This implies that the network can learn from raw sequences of different sizes.

4) CNN USED

The transformed skeleton map is used as input. We use an existing CNN with pre-trained weights on ImageNet as we find this ameliorates the classification score even when the images are handcrafted. We choose an 18-layer ResNet [12] for its compromise between accuracy and speed.

We extract a pose feature vector $\mathbf{s}$ from the skeleton map $\mathbf{M}$ with the pose module f_S with parameters θ_S in Equation 3. Here, and for the rest of the paper, subscripts of modules and parameters refer to a module, not an index.

$$\mathbf{s} = f_S(\mathbf{M}|\theta_S) \tag{3}$$

B. IR MODULE

The action performed by a subject is only a small region inside the frames of an IR sequence. The 2D skeleton data are used to capture the region of interest and virtually focus the attention of the network, with multiple potential subjects in mind. Because the IR module requires a video input with a fixed number of frames, a subsampling strategy is deployed. A 3D CNN is used to exploit the IR data.

1) CROPPING STRATEGY

Traditionally, 3D CNNs require a lot of parameters to account for the complex task of video understanding. Thus, the frames are heavily downscaled to reduce memory needs. In the process, discriminating information may be lost. In an action video of daily activities, the background provides little to no context. We would like our model to only focus on the subject as this is where the action happens. We argue that a crop around the subject provides ample cues about the action performed. Depth information, coupled with pose estimation algorithms, provides a turnkey solution for human detection. We propose a cropping strategy, shown Fig. 5 by a green parallelepiped, to virtually force the model to focus on the subject.

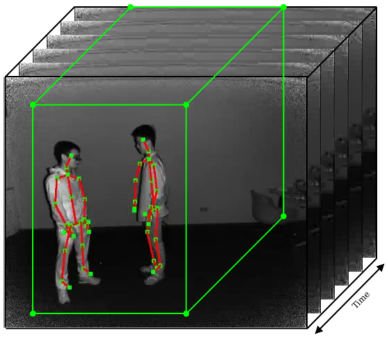

FIGURE 5. A fixed bounding box across the entire sequence is generated using the 2D skeleton information. The new sequence focuses attention on the subject rather than the background which provides little to no context. The images are taken from the NTU RGB+D dataset [38].

Given a 3D skeleton sequence projected on the 2D frames of the IR stream, we extract the maximal and minimal pixel positions across all joints and frames. This creates a fixed bounding box capturing the subject on the spatial and temporal domains. We empirically choose a 20 pixels offset to account for potential skeleton inaccuracy. The IR stream is padded with zeros should the box coordinates with the offset exceed the IR frame range.

The advantage of our method is as follows. Providing a crop around the region of interest reduces the size of the frames without decreasing the quality. The downscaling factor is thus less important and preserves a better aspect of the image. Furthermore, it alleviates the need for an attention mechanism as the cropping strategy may be seen as a hard attention scheme in itself. Also, the network does not have to learn information from the background, which is noise in our case, as it is reduced to a minimum.

2) MULTI-SUBJECT STRATEGY

The cropping strategy can be generalized to multiple subjects. The bounding box is enlarged to account for the other subjects. We take the maximal and minimal values across all joints, frames, and subjects.

For a given sequence, the bounding box is immobile regardless of the number of subjects. This allows keeping camera dynamics. We do not want to add confusion to the sequence by adding a virtual movement of the camera with a mobile bounding box.

3) SAMPLING STRATEGY

Contrary to the pose network, a given IR sequence is not treated in its entirety. A 3D CNN requires a sequence with a fixed number of frames T. Choices must be made regarding the value of T and the sampling strategy. A potential approach would be to take adjacent frames in a sequence. But the subsequence might not be enough to correctly capture the essence of the action. Instead, we propose a scheme where the raw sequence is divided into T windows of equal duration similar to [32], as illustrated Fig. 6. A random frame is taken from each window. A new sequence is created of length T.

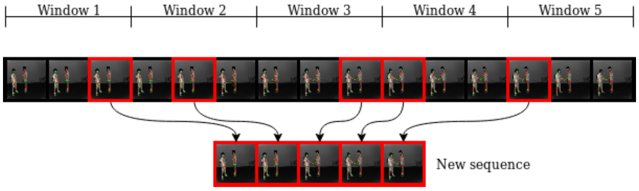

FIGURE 6. Each IR sequence is divided into a fixed number of windows of equal size. For each subdivision, a random frame is sampled. The concatenation of those frames is the input for the IR module.

4) 3D CNN USED

The new sampled sequences are used as inputs for the 3D CNN. We use an 18-layer deep R(2+1)D network [50] pre-trained on Kinetics-400 [3]. R(2+1)D is an elegant network which revisits 3D convolutions. Tran *et al.* showed factoring spatial and temporal convolutions yields state-of-the-art results on benchmark RGB action recognition datasets. Separating spatial and temporal convolutions with a nonlinear activation function in between allows for a more complex function representation with the same number of parameters.

We extract a stream feature vector $\mathbf{i}$ from the sampled IR sequence $\mathbf{I}$ with the IR module f_{IR} with parameters θ_{IR}, as follows:

$$\mathbf{i} = f_{IR}(\mathbf{I}|\theta_{IR}). \tag{4}$$

C. STREAM FUSION

Both pose and IR modules output their feature vectors. An MLP serves as the final module and returns a probability distribution for each action class in a dataset.

Features of both streams are fused using different schemes (average, sum, multiplication, max, convolution, concatenation). The MLP consists of three layers with batch normalization [17] before computation. The *ReLU* activation function is used for all neurons. Lastly, a *softmax* activation function is deployed to normalize the last layer's output into a probability distribution.

The class probability distribution $\mathbf{y}$ is outputted by the MLP f_{MLP} with parameters θ_{MLP} in Equation 5. Inputs $\mathbf{i}$ and $\mathbf{s}$ correspond to the feature vectors computed by the pose and IR modules.

$$\mathbf{y} = f_{MLP}(\mathbf{i}, \mathbf{s}|\theta_{MLP}) \tag{5}$$

We tried a scheme where the pose and IR modules of our network would emit their own prediction. We would then average the predictions on a logits level with learned weights during the backpropagation step. However, this would lead to the network's final classification to be attributed solely to one module or the other. Instead, we believe that an MLP allows for the features of the two streams to be interpreted jointly.

IV. NETWORK ARCHITECTURE

A. ARCHITECTURE

1) POSE MODULE

The pose network is an 18-layer deep ResNet [12]. The network takes as input a tensor of dimensions 3 × 224x224, where 3 corresponds to the RGB channels and 224 to the height and width of the image. The output, $\mathbf{s}$, is a 1D vector of 512 features.

2) IR MODULE

The IR network is an 18-layer deep R(2+1)D [50]. It takes as input a video of dimensions 3xTx112 × 112, where 3 corresponds to the RGB channels, T to the length of the sequence, and 112 to the height and width of the image. The output, $\mathbf{i}$, is a 1D vector of 512 features.

To be able to leverage the pre-trained R(2+1)D CNN, which is originally trained on RGB images, the IR frames, which are single-channel grayscale images, are duplicated.

3) CLASSIFICATION MODULE

The classification module is an MLP network with three layers. The first layer expects a vector of 512 (average, sum, multiplication, max, convolution) or 1024 (concatenation) features and comprises 256 units. The second layer consists of 128 units. The last layer has as many units as there are different action classes in a dataset. Finally, the *softmax* function is used to normalize the predictions to a probability distribution. Batch normalization is applied before the layers. A dropout scheme has been tested in place of batch normalization but was not found to be superior.

The entire network, detailed Fig. 7, is trained in end-to-end fashion. The weights are reset after each run.

B. DATA AUGMENTATION

To prevent overfitting and reinforce the generalization capabilities of our model, we perform data augmentation during training.

The skeleton sequences have limited viewpoints but their representation makes them excellent candidates for augmentation through geometric transformations. The skeleton

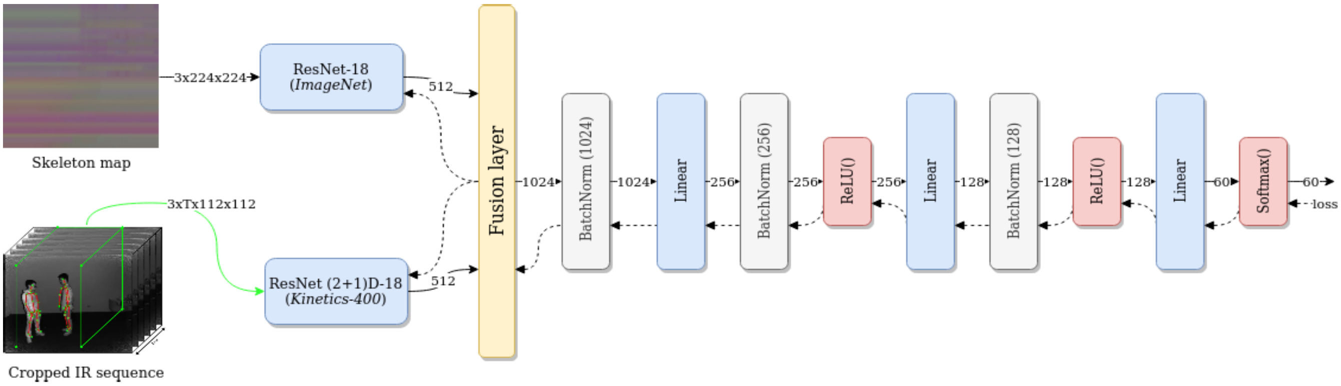

FIGURE 7. **The full detailed model. The pose and IR modules output separate feature vectors. The two are fused and a final MLP outputs a class probability distribution. The pose network is a pre-trained *ResNet-18*. The IR network is a pre-trained *R(2+1)D-18* network.**

sequences are enhanced by performing a random rotation around the X, Y and Z axis. For each sequence during training, we apply a random rotation between $-20°$ and $20°$ on each axis.

We approach IR data augmentation with the following scheme. For each sequence during training, we perform a horizontal mirroring transformation on the frames with a 50% chance probability.

When training the entire FUSION model, the best results are achieved when the two streams are augmented independently compared to mirroring the skeleton data (e.g. the left hand becomes the right hand) jointly with the IR video.

C. TRAINING

The network is trained in end-to-end fashion by minimizing cross-entropy loss, meaning all the modules of our network are trained together. The pose network is pre-trained on the ImageNet dataset [6]. The IR network is pre-trained on the Kinetics-400 dataset [3].

V. EXPERIMENTS

We evaluate the performances of our proposed model on the NTU RGB+D dataset, the largest benchmark to date [38]. We also perform extensive ablation studies to understand the individual contributions of our modules.

A. NTU RGB+D DATASET

The NTU RGB+D dataset [38] is the largest human action recognition dataset to date captured with a Microsoft Kinect V2 [66]. To our knowledge, it is also the only one including the IR sequences. It contains 60 different classes ranging from daily to health-related actions spread across 56,880 clips and 40 subjects. It includes 80 different views. An action may require up to two subjects. The skeletons are composed of 25 joints and the IR videos are all in 512×424. The various setups, views, orientations, result in a great diversity which makes NTU RGB+D a challenging dataset.

We follow the two official benchmark evaluations for this dataset: Cross-Subject (CS) and Cross-View (CV). The former splits the 40 subjects into training and testing groups so that all sequences of a subject are in one set or the other. The latter uses the samples acquired from cameras 2 and 3 for training while the samples from camera 1 are used for testing.

B. EXPERIMENTAL SETTINGS

For consistency, we do not modify the following hyperparameters across all experiments. We set the batch size to 16. Gradient clipping is used to avoid an exploding gradient issue. We set it to 10. Adam optimizer [24] is used to train the networks. A learning rate of 0.0001 is set and kept consistent during training.

The pose and IR modules each require a fixed input size. Skeleton maps are resized to 224×224 images. IR frames are resized to 112×112.

To assure consistency and reproducibility, we use a pseudorandom number generator fed with a fixed seed. Following [38], we sample 5% of the training set as our validation set. The model with the highest validation accuracy is used to evaluate the test set. We perform each experiment five times, with a different seed, for a total of 10 runs across the two benchmarks. Mean and standard deviation of the accuracy scores are reported. Statistical significance is evaluated using a one-tailed paired sample T-Test. Experiments are paired by benchmark and seed. For example, when evaluating data augmentation, the accuracy of the model without augmentation is paired with its augmented counterpart for the same seed. As such, each sample contains ten results. A critical value of 0.05 is used. Unless specified otherwise, a discussed increase in performance is found statistically significant.

C. ABLATION STUDIES

In this section, we isolate the pose and IR modules and study different parameters. Action classification accuracy on the NTU RGB+D dataset is used as the comparison metric. We evaluate the impact of various fusion schemes, transfer learning, data augmentation, pose conditioning of IR sequences, and the number of frames T. Finally, we compare our results with the current state of the art.

The CV benchmark is a much easier task. The test actions are already seen during training but from a different point of view with a different camera. Although the different setups yield different joint position estimations for a given sequence [65], the geometric nature of skeleton data allows for a better generalization. This is not the case for the CS task as the test sequences are completely novel. Consequently, the following discussions will only address the CS benchmark.

1) POSE MODULE ALONE

We evaluate the performances of our pose module as a standalone. The IR module does not intervene. We also adjust the input size of the classification MLP. Optimal results are achieved by combining pre-training with data augmentation. Table 7 shows the best results of the pose module on NTU RGB+D: 81.9% on CS and 89.6% on CV.

2) INFRARED MODULE ALONE

The other part of the FUSION network, and arguably the most important contributor, is the infrared module. In a similar fashion as above, the input size of the MLP is adjusted while keeping the number of neurons equal. Optimal results are achieved with a pre-trained network, with data augmentation, on pose-conditioned inputs for a sequence length of $T = 20$. Table 7 shows the performance of the IR module as a standalone: 90.4±0.79% on CS and 93.8±0.46% on CV.

3) INFLUENCE OF FEATURE FUSION SCHEME

We test various deep feature fusion schemes: average (avg), sum, multiplication (mult), max, convolution (conv) and concatenation (concat). The average, sum, multiplication, and max fusion schemes are done in element-wise fashion. The convolution scheme considers the **i** and **s** feature vectors as a 512×2 image. The features are convoluted by a 2D kernel of size $(1, 2)$. A new 1D feature vector with 512 new computed features is thus outputted. Table 1 shows the impact of the different fusion schemes on classification accuracy for the CS and CV benchmarks.

TABLE 1. Impact of fusion scheme on classification performances (A: Augmented | P: Pre-trained | C: cropped inputs) (accuracy in %).

Method	Pose	IR	CS	CV	Average
FUSION (conv) - CPA	X	X	91.3±0.47	94.2±0.56	92.75
FUSION (mult) - CPA	X	X	91.4±0.38	94.1±0.26	92.75
FUSION (concat) - CPA	X	X	91.3±0.25	94.5±0.31	92.90
FUSION (max) - CPA	X	X	**91.8±0.37**	94.5±0.31	93.15
FUSION (avg) - CPA	X	X	91.6±0.42	94.7±0.25	93.15
FUSION (sum) - CPA	X	X	**91.8±0.40**	**94.9±0.39**	**93.35**

The different schemes perform similarly. More convincingly, the sum scheme (93.35% average on CS and CV) has the highest mean accuracy, but not found to be statistically better than the average and max schemes. Nonetheless, regardless of the chosen scheme, results are systematically improved compared to the pose and IR modules (Table 7).

4) INFLUENCE OF PRE-TRAINING

Pre-training a network is an elegant way to transfer a learned task to a new one. It has been shown to provide impressive results even on handcrafted images [65]. Furthermore, it helps with the overfitting issue smaller datasets may demonstrate.

We evaluate the impact of this strategy on our network. Table 2 shows the effect of pre-training on the different modules.

TABLE 2. Impact of pre-training on classification performances (P: Pre-trained | C: cropped inputs) (accuracy in %).

Method	Pose	IR	CS	CV
Pose module	X	-	78.0±0.49	84.7±0.58
Pose module - P	X	-	**81.0±0.48**	**87.4±0.45**
IR module	-	X	75.8±0.60	74.2±2.10
IR module - P	-	X	**83.7±0.70**	**85.5±0.64**
IR module - C	-	X	84.2±0.32	88.6±0.33
IR module - CP	-	X	**90.1±0.39**	**92.7±0.95**

The pose network enjoys a noticeable increase in accuracy of about 2.5% for both benchmarks (78.0% to 81.0% on CS). It is pre-trained on ImageNet, which consists of real-life images. The skeleton maps used as inputs are handcrafted. Even then, a pre-training scheme shows encouraging results.

The impact of pre-training on the IR module's accuracy is significant. For uncropped sequences, the accuracy increases by about 8% for both benchmarks (75.8% to 83.7% on CS). For cropped sequences, the gain is almost 6% for the cross-subject benchmark (84.2% to 90.1%) and above 4% for cross-view (88.6% to 92.7%).

The greater contribution of transfer learning for the IR module compared to the pose module might be explained by the greater resemblance of IR vs. RGB videos compared to handcrafted vs. real-life images. Nonetheless, such findings further emphasize the power of transfer learning.

5) INFLUENCE OF DATA AUGMENTATION

Data augmentation consists of virtually enlarging the dataset, thus hopefully preventing overfitting and reducing variance between training and test sets. We perform augmentation for the different streams. Table 3 shows the performances of data augmentation on the different modules with pre-trained networks. Overall, data augmentation yields favorable results.

The pose module alone enjoys an increase of about 1% accuracy for CS and 1.5% for CV (81.0% to 81.9% on CS). Mirroring skeleton data with a 50% chance during training, in addition to random rotations further improves the result, especially on CV. The IR module alone seems to benefit more from data augmentation on the CV benchmark compared to the CS. For the CV benchmark, the increase is about 1% whether the input sequence is cropped (92.7% to 93.8%) or not (85.5% to 86.7%). Overall, the improvements are significant. When the modules are fused using an element-wise sum scheme, our FUSION network, independent data augmentation is favorable with an increase of 0.7% for the CS benchmark (91.1% to 91.8%) and 0.8% for the CV benchmark (94.1% to 94.9%). Mirroring the skeleton and IR data

TABLE 3. Impact of data augmentation on classification performances (A: Augmented | P: Pre-trained | C: cropped inputs | †: augmentation is joint, i.e. when an IR sequence is mirrored, so is the skeleton) (accuracy in %).

Method	Pose	IR	CS	CV
Pose module - P	X	-	81.0±0.48	87.4±0.45
Pose module - PA (rot)	X	-	81.9±0.82	89.0±0.39
Pose module - PA (rot + mirror)	X	-	**81.9±0.28**	**89.6±0.53**
IR module - P	-	X	83.7±0.70	85.5±0.64
IR module - PA	-	X	**84.5±0.84**	**86.7±0.89**
IR module CP	-	X	90.1±0.39	92.7±0.95
IR module CPA	-	X	**90.4±0.79**	**93.8±0.46**
FUSION - CP (sum)	X	X	91.1±0.21	94.1±0.45
FUSION - CPA (sum)	X	X	**91.8±0.40**	**94.9±0.39**
FUSION - CPA (sum) †	X	X	91.6±0.37	94.2±0.44

TABLE 4. Impact of our cropping strategy on classification performances (A: Augmented | P: Pre-trained | C: cropped inputs) (accuracy in %).

Method	Pose	IR	CS	CV
IR module (baseline)	-	X	75.8±0.60	74.2±2.10
IR module- PA	-	X	**84.5±0.84**	**86.7±0.89**
IR module - C	-	X	84.2±0.32	88.6±0.33
IR module - CPA	-	X	**90.4±0.79**	**93.8±0.46**

jointly also leads to improved performances, but to a lesser extent.

6) TRANSFER LEARNING VS. DATA AUGMENTATION

Transfer learning and data augmentation are two strategies to better generalize the performances of a network. Transfer learning leverages the learned parameters from another dataset while data augmentation virtually enlarges the current dataset. A small dataset might lead to overfitting which increases variance between the training and validation sets as the training error continues to lower.

Our model can reach a negligible training error, even with individual modules, showcasing an overfitting issue. Having studied the impacts on performances of both methods, transfer learning shows much better results. This might be explained by the already large size of the NTU RGB+D dataset mitigating the potential of data augmentation. Nonetheless, it is formidable how a model can yield vastly different performances based on the initialization of its parameters. The black-box nature of deep learning makes the interpretation of how a model learns difficult. Perhaps future works will focus on understanding the internal representation of a network to guide its learning rather than implementing evermore complex models.

7) INFLUENCE OF POSE-CONDITIONED CROPPED IR SEQUENCES

In this section, we evaluate the impact of our cropping strategy, detailed section III-B1, on the performances of the IR module as a standalone. Table 4 shows a significant increase in performances.

Our baseline for this comparison, the IR module without transfer learning and data augmentation on uncropped sequences, reports unsatisfactory results (75.8% on CS). With transfer learning and data augmentation, we are able to increase the accuracy by 10% average for both benchmarks (75.8% to 84.5% on CS). However, we find that our cropping strategy alone reaps similar benefits (75.8% to 84.2% on CS). When combining all three strategies, we further ameliorate the classification score by about 5% (90.4% on CS). The average gain for both benchmarks is thus above 15%, which is considerable.

We demonstrate the power of a pragmatic approach. An identical model performs significantly better thanks to careful design choices.

8) INFLUENCE OF SEQUENCE LENGTH

Sequences of the NTU RGB+D dataset are at most a couple of seconds long. We study the impact of the length T of the new sampled IR sequence on classification performances of two networks: the IR module only and on the complete FUSION model. Both models are pre-trained and fed with augmented data. The IR sequences are pose-conditioned. Table 5 reports the impact of different values of T on the accuracy score.

As a general tendency, the greater the value of T, the better the results. Best results are achieved for $T = 20$ (on CS: 90.4% for IR module only and 91.8% for FUSION). For the FUSION network, excellent results are achieved for a number of frames as little as $T = 8$ (89.5% on CS and 92.9% on CV). The differences between $T = 12$ and $T = 16$ are not significant. The results really shine with $T = 20$. But FUSION networks with a smaller value of T are much faster, showcasing a trade-off between speed and accuracy.

9) PERFORMANCES BASED ON ACTION TYPE

We separate the action classes into three categories: intense kinetic movement, similar motion, and object-related actions. Details are provided in the footnote page 10. We use the class IDs defined in [38]. The pose module, the IR module, and the entire FUSION model are evaluated Table 6. For more details on single-class performances, confusion matrices for all modules can be found on the project page.

The pose module has a strong ability to classify actions with intense movements (86.3% on CS) compared to similar motion (76.1%) and object-related actions (76.8%). Actions such as sitting down, standing up, falling, jumping, staggering, walking toward or away from another subject are classified with over 95% accuracy. Unsurprisingly, similar motion and object-related actions prove the most challenging. Writing is the trickiest, with 40% accuracy only and often mislabeled as writing or typing on a keyboard. We believe this will always be a limitation of pose-only networks.

The IR module has a more balanced accuracy score. Some actions, such as touching another person's pocket or staggering, prove more difficult to recognize for the IR module compared to the pose module. However, some object-oriented actions are still difficult to correctly discern. For instance, writing is more often than not mislabeled as playing with a phone. We propose two possible explanations. Firstly, the object information might be lost during the rescaling process, even with our cropping strategy in place. Secondly, the IR

TABLE 5. Impact of IR sequence length on classification performances (A: Augmented | P: Pre-trained | C: cropped inputs) (accuracy in %).

Method	Pose	IR	CS				CV			
			T=8	T=12	T=16	T=20	T=8	T=12	T=16	T=20
IR module - CPA	-	X	86.6±0.27	88.9±0.50	89.8±0.45	**90.4±0.79**	89.0±0.39	91.5±0.65	93.0±0.37	**93.8±0.46**
FUSION - CPA (sum)	X	X	89.5±0.41	90.2±0.30	90.5±0.68	**91.8±0.40**	92.9±0.56	93.6±0.54	93.7±0.47	**94.9±0.39**

TABLE 6. Comparison by action category. Class IDs can be found on the NTU-RBG+D website. (A: Augmented | P: Pre-trained | C: cropped inputs) (accuracy in %).

Method	Pose	IR	CS			CV		
			Intense movement[2]	Similar motion[3]	Object-related[4]	Intense movement	Similar motion	Object-related
Pose module	X	-	86.3±0.32	76.1±0.82	76.8±0.78	92.1±0.37	84.0±0.89	85.5±1
IR module - CPA	-	X	91.6±0.44	88.7±0.93	89.6±1.16	94.2±0.61	93.1±0.68	93.9±0.77
FUSION - CPA (sum)	X	X	**93.1±0.53**	**90.0±0.92**	**90.5±0.39**	**95.6±0.2**	**93.3±0.8**	**94.3±0.72**

TABLE 7. Comparison of our model to the state of the art (A: Augmented | P: Pre-trained | C: cropped inputs) (accuracy in %).

Method	Pose	RGB	Depth	IR	CS	CV
Lie Group [53]	X	-	-	-	50.1	82.8
HBRNN [9]	X	-	-	-	59.1	64.0
Deep LSTM [38]	X	-	-	-	60.7	67.3
PA-LSTM [38]	X	-	-	-	62.9	70.3
ST-LSTM [32]	X	-	-	-	69.2	77.7
STA-LSTM [47]	X	-	-	-	73.4	81.2
VA-LSTM [64]	X	-	-	-	79.2	87.7
TCN [23]	X	-	-	-	74.3	83.1
C+CNN+MTLN [21]	X	-	-	-	79.6	84.8
Synth. CNN [33]	X	-	-	-	80.0	87.2
3scale ResNet [27]	X	-	-	-	85.0	92.3
DSSCA-SSLM [39]	-	X	X	-	74.9	-
[36]	X	-	X	-	75.2	83.1
CMSN [67]	X	X	-	-	80.8	-
STA-HANDS [2]	X	X	-	-	84.8	90.6
Coop CNN [58]	-	X	X	-	86.4	89.0
ST-GCN [62]	X	-	-	-	81.5	88.3
DGNN [41]	X	-	-	-	89.9	**96.1**
Pose module - PA	X	-	-	-	81.9±0.28	89.6±0.53
IR module - CPA	-	-	-	X	90.4±0.79	93.8±0.46
FUSION - CPA (sum)	X	-	-	X	**91.8±0.40**	**94.9±0.39**

nature, grayscale and noisy, might not be clear enough to discern the object correctly. But other object-related actions such as dropping an object or brushing hair see an impressive improvement of over 10%.

The FUSION network is able to benefit from the kinetic information of the pose module to improve the accuracy of actions difficult for the IR module. For instance, touching the neck is improved from 82% (pose only) and 77% (IR only) to 95%. This is a strong demonstration that the two feature networks work conjointly. However, the FUSION network is still challenged by hand-related actions (eat a meal, brush teeth, reading, writing).

10) COMPARISON WITH THE STATE OF THE ART

We compare our FUSION model, using an average fusion scheme, with the state of the art (Table 7). We divide current methods into 5 different frameworks including handcrafted features, RNN-based methods, CNN-based methods, fusion methods, and GCN-based methods. Current best results are obtained using skeleton data only with GCNs. We achieve better results than the current state of the art on the CS benchmark (91.8%) with 1.9% accuracy increase. On the CV benchmark, results are comparable (91.8±0.40% for FUSION against 91.8±0.40% for DGNN [41]). We conclude to the efficacy of IR data to correctly interpret human actions. Given the representation power of IR, follow-up works should compare this stream to RGB and/or depth maps.

We significantly improve upon current fusion methods, once again validating the complementary role of pose and visual data.

VI. CONCLUSION

We propose an end-to-end trainable network using skeleton and infrared data for human action recognition. The pose and infrared modules report strong individual performances, which is greatly due to the power of transfer learning as they are both pre-trained on other large scale datasets. When working in symbiosis, the results are further ameliorated. We are the first to jointly use pose and infrared streams. Our method improves the state of the art on the largest RGB-D action recognition dataset to date. Compared to other fusion approaches, our method uses a single video stream, which we believe is more memory efficient.

Our work demonstrates the strong representational power of infrared data, which opens the door for applications where illumination conditions render RGB videos unusable. The complementary role of pose and visual streams is further illustrated, which is in line with previous work. Given the limits of our network on hand-related actions, future work could focus on integrating a dedicated stream.

ACKNOWLEDGMENT

The authors would also like to thank their collaborators from Aerosystems International Inc.

REFERENCES

[1] J. K. Aggarwal and L. Xia, "Human activity recognition from 3D data: A review," *Pattern Recognit. Lett.*, vol. 48, pp. 70–80, Oct. 2014.

[2] F. Baradel, C. Wolf, and J. Mille, "Pose-conditioned spatio-temporal attention for human action recognition," 2017, *arXiv:1703.10106*. [Online]. Available: http://arxiv.org/abs/1703.10106

[2] Class IDs: 7, 8, 9, 10, 22, 23, 24, 26, 27, 31, 34, 36, 37, 38, 40, 42, 43, 44, 45, 46, 47, 48, 50, 51, 52,53, 54, 55, 56, 57, 58, 59, 60

[3] Class IDs: 1, 2, 11, 12, 14, 15, 16, 17, 18, 19, 20, 21, 29, 30, 34, 35, 36, 39

[4] Class IDs: 1, 2, 3, 4, 5, 6, 11, 12, 13, 14, 15,16, 17, 18, 19, 20, 21, 28, 29, 30, 32, 33, 49

[3] 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.

[4] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1875–1886, Nov. 2015.

[5] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "MARS: Motion-augmented RGB stream for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7882–7891.

[6] 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.

[7] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, T. Darrell, and K. Saenko, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.

[8] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," in *Proc. 3rd IAPR Asian Conf. Pattern Recognit. (ACPR)*, Nov. 2015, pp. 579–583.

[9] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1110–1118.

[10] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1933–1941.

[11] F. Han, B. Reily, W. Hoff, and H. Zhang, "Space-time representation of people based on 3D skeletal data: A review," *Comput. Vis. Image Understand.*, vol. 158, pp. 85–105, May 2017.

[12] 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.

[13] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 3, pp. 807–811, Mar. 2018.

[14] Y. Hou, S. Wang, P. Wang, Z. Gao, and W. Li, "Spatially and temporally structured global to local aggregation of dynamic depth information for action recognition," *IEEE Access*, vol. 6, pp. 2206–2219, 2018.

[15] J.-F. Hu, W.-S. Zheng, J. Pan, J. Lai, and J. Zhang, "Deep bilinear learning for RGB-D action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 335–351.

[16] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3D joint locations," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 1–7.

[17] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[18] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Perception Psychophys.*, vol. 14, no. 2, pp. 201–211, Jun. 1973.

[19] 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.

[20] Q. Ke, S. An, M. Bennamoun, F. Sohel, and F. Boussaid, "SkeletonNet: Mining deep part features for 3-D action recognition," *IEEE Signal Process. Lett.*, vol. 24, no. 6, pp. 731–735, Jun. 2017.

[21] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid, "A new representation of skeleton sequences for 3D action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3288–3297.

[22] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel(R) realsense(TM) stereoscopic depth cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1–10.

[23] T. S. Kim and A. Reiter, "Interpretable 3D human action analysis with temporal convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 1623–1631.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[25] 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.*, Nov. 2011, pp. 2556–2563.

[26] I. Lee, D. Kim, S. Kang, and S. Lee, "Ensemble deep learning for skeleton-based action recognition using temporal sliding LSTM networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1012–1020.

[27] B. Li, Y. Dai, X. Cheng, H. Chen, Y. Lin, and M. He, "Skeleton based action recognition using translation-scale invariant image mapping and multi-scale deep CNN," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2017, pp. 601–604.

[28] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance maps based action recognition with convolutional neural networks," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 624–628, May 2017.

[29] C. Li, Y. Hou, P. Wang, and W. Li, "Multiview-based 3-D action recognition using deep networks," *IEEE Trans. Human-Machine Syst.*, vol. 49, no. 1, pp. 95–104, Feb. 2019.

[30] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3595–3603.

[31] C. Liu, Y. Hu, Y. Li, S. Song, and J. Liu, "PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding," 2017, *arXiv:1703.07475*. [Online]. Available: http://arxiv.org/abs/1703.07475

[32] J. Liu, A. Shahroudy, D. Xu, and G. Wang, "Spatio-temporal LSTM with trust gates for 3D human action recognition," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Oct. 2016, pp. 816–833. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46487-9_50

[33] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognit.*, vol. 68, pp. 346–362, Aug. 2017.

[34] V. Mnih, N. Heess, and A. Graves, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[35] L. Lo Presti and M. La Cascia, "3D skeleton-based human action classification: A survey," *Pattern Recognit.*, vol. 53, pp. 130–147, May 2016.

[36] H. Rahmani and M. Bennamoun, "Learning action recognition model from depth and skeleton videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5832–5841.

[37] 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.

[38] 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. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.

[39] A. Shahroudy, T.-T. Ng, Y. Gong, and G. Wang, "Deep multimodal feature analysis for action recognition in RGB+D videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1045–1058, May 2018.

[40] S. Sharma, R. Kiros, and R. Salakhutdinov, "Action recognition using visual attention," 2015, *arXiv:1511.04119*. [Online]. Available: http://arxiv.org/abs/1511.04119

[41] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Skeleton-based action recognition with directed graph neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7912–7921.

[42] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12026–12035.

[43] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *Proc. CVPR*, Jun. 2011, pp. 1297–1304.

[44] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan, "An attention enhanced graph convolutional LSTM network for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1227–1236.

[45] M. Simonovsky and N. Komodakis, "Dynamic edge-conditioned filters in convolutional neural networks on graphs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3693–3702.

[46] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[47] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1–8.

[48] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*. [Online]. Available: http://arxiv.org/abs/1212.0402

[49] 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.

[50] 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.

[51] Z. Tu, H. Li, D. Zhang, J. Dauwels, B. Li, and J. Yuan, "Action-stage emphasized spatiotemporal VLAD for video action recognition," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2799–2812, Jun. 2019.

[52] Z. Tu, W. Xie, Q. Qin, R. Poppe, R. C. Veltkamp, B. Li, and J. Yuan, "Multi-stream CNN: Learning representations based on human-related regions for action recognition," *Pattern Recognit.*, vol. 79, pp. 32–43, Jul. 2018.

[53] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a lie group," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 588–595.

[54] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3551–3558.

[55] H. Wang and L. Wang, "Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 499–508.

[56] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1290–1297.

[57] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "RGB-D-based human motion recognition with deep learning: A survey," *Comput. Vis. Image Understand.*, vol. 171, pp. 118–139, Jun. 2018.

[58] P. Wang, W. Li, J. Wan, P. Ogunbona, and X. Liu, "Cooperative training of deep aggregation networks for RGB-D action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–8.

[59] P. Wang, Z. Li, Y. Hou, and W. Li, "Action recognition based on joint trajectory maps using convolutional neural networks," in *Proc. ACM Multimedia Conf. (MM)*, 2016, pp. 102–106.

[60] Y. Xiao, J. Chen, Y. Wang, Z. Cao, J. T. Zhou, and X. Bai, "Action recognition for depth video using multi-view dynamic images," *Inf. Sci.*, vol. 480, pp. 287–304, Apr. 2019.

[61] K. Xu, W. Hu, J. Leskovec, and S. Jegelka, "How powerful are graph neural networks?" 2018, *arXiv:1810.00826*. [Online]. Available: http://arxiv.org/abs/1810.00826

[62] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–11.

[63] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, "Describing videos by exploiting temporal structure," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4507–4515.

[64] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2117–2126.

[65] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive neural networks for high performance skeleton-based human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1963–1978, Aug. 2019.

[66] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE Multimedia Mag.*, vol. 19, no. 2, pp. 4–10, Feb. 2012.

[67] M. Zolfaghari, G. L. Oliveira, N. Sedaghat, and T. Brox, "Chained multi-stream networks exploiting pose, motion, and appearance for action classification and detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2904–2913.

ALBAN MAIN DE BOISSIERE is currently pursuing the M.Eng. degree with the Institut National des Sciences Appliquées (INSA), Lyon, and the M.A.Sc. degree with the École de Technologie Supérieure (ETS), Montreal. His major interests include computer vision problems related to action recognition, early prediction, and online action detection.

RITA NOUMEIR (Member, IEEE) received the master's and Ph.D. degrees in biomedical engineering from École Polytechnique of Montreal. She is currently a Full Professor with the Department of Electrical Engineering, École de Technologie Superieure (ETS), Montreal. Her main research interest is in applying artificial intelligence methods to create decision support systems. She has extensively worked in healthcare information technology and image processing. She has also provided consulting services in large-scale software architecture, healthcare interoperability, workflow analysis, technology assessment, and image processing for several international software and medical companies, including Canada Health Infoway.