2022 2nd International Conference on Computing and Information Technology (ICCIT)
Jan. 25 - 27, 2022/ FCIT/UT/KSA

An Attention-based Hybrid 2D/3D CNN-LSTM for Human Action Recognition

Khaled Bayoudh
Electrical Department, National Engineering School of Monastir (ENIM), Laboratory of Electronics and Micro-electronics (LR99ES30), Faculty of Sciences of Monastir (FSM), University of Monastir, Monastir, Tunisia
khaled.isimm@gmail.com

Fayçal Hamdaoui
Electrical Department, National Engineering School of Monastir (ENIM), Laboratory of Control, Electrical Systems and Environment (LASEE), National Engineering School of Monastir (ENIM), University of Monastir, Monastir, Tunisia
faycel_hamdaoui@yahoo.fr

Abdellatif Mtibaa
Electrical Department, National Engineering School of Monastir (ENIM), Laboratory of Electronics and Micro-electronics (LR99ES30), Faculty of Sciences of Monastir (FSM), University of Monastir, Monastir, Tunisia
abdellatif.mtibaa@enim.rnu.tn

2022 2nd International Conference on Computing and Information Technology (ICCIT) | 978-1-6654-3605-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICCIT52419.2022.9711631

***Abstract*—Human Action Recognition (HAR) is a challenging problem in computer vision that has received a great deal of attention in the last decade. With the advent of new deep learning techniques such as convolutional neural networks (CNNs), the recognition performance of HAR systems has improved significantly over traditional methods, mainly due to the powerful representation capabilities of CNNs. In most of the literature, 2D CNNs or their 3D counterparts have been used to learn spatial and temporal image-level features of videos. In this paper, we developed an end-to-end HAR framework based on a hybrid 2D/3D CNN. The hybrid CNN feature extractor aims to exploit the potential collaboration between 2D and 3D CNNs. The CNN features extracted from the video sequences are then fed into a Long Short-Term Memory (LSTM) network to capture the short- and long-term dependencies in the data structure. Inspired by human visual attention mechanisms, a visual attention module was used in this study to focus semantically on relevant salient features in visual representations. The developed model was trained and evaluated using the KTH dataset and achieved promising recognition performance compared to state-of-the-art methods.**

Keywords—Attention mechanisms, Human action recognition, Hybrid 2D/3D CNN, LSTM

I. Introduction

Recognizing human behavior is a core problem of computer vision that has been widely studied in a wide range of practice areas. In particular, video-based human action recognition (HAR) [1] tasks have received a lot of attention from the deep learning community in the last few years due to their ability to support many cutting-edge applications and real-time embedded systems, such as healthcare, smart video surveillance, human-computer interaction, autonomous driving systems, and so on. From a practical standpoint, the main task of HAR is to discriminate between different types of actions in video clips captured by image sensors, mounted on mobile or fixed devices. In other words, it is a problem of predicting the actions in a video for a short-term period. The recognition system first detects intrinsic patterns of a video sequence consisting of multiple frames and then attempts to predict an activity or action of interest on a spatio-temporal axis. Since their introduction by LeCun et al. [2], deep learning algorithms have made remarkable progress in the field of pattern recognition and machine learning due to their superior performance and versatility. In the current literature, HARs can be divided into two categories: hand-crafted learning-based methods and deep learning-based methods. Hand-crafted learning-based methods are based on a combination of feature and descriptor extractors (e.g., HOG, SIFT, etc.) and shallow classifiers (SVM, random forest, etc). Indeed, the advent of deep learning techniques has greatly improved the generic task of object recognition in terms of accuracy and efficiency, but there are still several scenarios and situations that pose significant challenges in the development of HAR systems, such as geometric distortions, partial or complete occlusions, background clutter, high-speed motion, etc. Moreover, deep learning algorithms require a large amount of training data to achieve a reasonable level of prediction performance.

Convolutional neural network (CNN) [3] is one of the feed-forward deep learning models that exhibit strong discriminative and feature representation performance among other discriminative networks. First, CNNs extract a set of feature maps from the input image data and then make inferences based on the visual appearance of these features. Inspired by the remarkable success of CNNs in various computer vision tasks such as image classification and object recognition, recent literature has leveraged the discriminative power of these models for action recognition in video. Recently, the use of deep CNN models to adaptively learn dynamic structures from multidimensional data cubes in an end-to-end manner provides an effective and efficient solution for action recognition. Most recent research has focused on extracting and learning spatial and temporal feature vectors using only 2D or 3D CNNs to recognize actions in videos [4]. Both versions of CNNs (i.e., 2D and 3D CNNs) are based on adaptive learning schemes on static and temporal data (which can be time-synchronized image frames) provided by various types of visual sensors. In practice, 2D CNNs take only 2D image shapes as input and attempt to extract spatial data features from each frame, regardless of temporal variation and contextual dimensionality. In contrast, 3D CNNs perform 3D convolution on 3D data points and have the potential to extract contextual correlations and visual patterns from a high-dimensional data space, which makes them more suitable than their 2D counterparts for many video analysis tasks, including HAR tasks. While 3D CNN models have the advantage of being able to recognize actions in different scenes and scenarios, they also have some inherent shortcomings. For example, from a practical perspective, the use of 3D CNNs is very limited in the current literature because they increase the computational cost of deep models compared to 2D CNNs and typically require more internal memory and more computational power for the underlying computations. To overcome these limitations, inspired by the work of [5], [6], we adaptively combined 2D and 3D CNNs into a hybrid architecture in this work, taking advantage of the computational sharing between multidimensional CNN networks. This strategy can stabilize the learning algorithm and enhance the discriminative and computational burden of the action recognition model.

However, effectively modeling spatio-temporal information is still a challenging problem for most video-based action recognition systems. In recent years, Recurrent Neural Networks (RNNs), especially Long-Short Term Memory (LSTMs) [7], have been mainly used to encode both short- and long-range dependencies between data features when the temporal dimension of the frame sequence is considered. Also, to improve the contextual capability and computational efficiency of predictive models, the deep learning community has proposed the use of visual attention mechanisms [8] as a way to further process spatio-temporal information. This can be achieved by semantically focusing on the relevant salient feature representations in visual data.

In this paper, we proposed an attention-based 2D/3D hybrid CNN-LSTM network architecture for action recognition based on adaptive hybridization of hierarchical networks and an attention mechanism. The proposed hybrid architecture is designed to model rich multi-domain features by increasing the contextual correlation of spatio-temporal cues and exploiting synergies and computational sharing among processing modules. Experimental results show that our hybrid network outperforms recent state-of-the-art models on KTH dataset.

II. Related Works

A. Handcrafted methods

For decades, HAR has been one of the challenging tasks in computer vision and has made tremendous progress in the last few years. HAR consists of interpreting human actions and assigning consistent labels with similar appearance properties to each action. In general, an action recognition pipeline consists of three main steps. First, a video dataset is captured from an image sensor and pre-processed to create a background model. Next, local and global descriptors can be used to extract low-level features. Finally, a shallow classifier is employed to learn the extracted features and classify potential actions. Traditionally, most action recognition systems were based on low-level, handcrafted patterns as feature vectors for recognizing human actions in videos. Most of these methods rely on modeling the spatial and temporal structure of a video to identify frame-level action patterns in each sequence. The temporal dimension refers to information about variations in the motion of pixel positions in successive frames on the time axis. Several algorithms, such as Bag-of-Words (BoW) [9], template matching [10], motion- and trajectory-based methods [11], [12], etc., were proposed to analyze the input video data and extract temporal points of interest and local descriptors. To perform the final classification of the extracted features, conventional machine learning algorithms such as decision trees [13], support vector machines (SVM) [14], Naive Bayes [15], K-nearest kneighbors (KNN) [16], etc., have been mainly used in the literature. Experimental analysis has shown the effectiveness of these algorithms in several circumstances and scenarios. However, these methods rely on very complex and time-consuming feature extraction schemes, and the pattern recognition ability depends on the quality of the descriptors and the visual complexity of scenes, which can lead to degradation of recognition performance. This is where the deep learning paradigm takes place.

B. Deep learning methods

Traditional handcrafted-based action recognition methods are time-consuming and labor-intensive and typically perform poorly on complex scenes and challenging scenarios, such as occlusions, background blur, visual appearance variations, etc., making human behavior recognition analysis extremely difficult. For this reason, the computer vision community has turned to deep learning algorithms to address most of these limitations. Given the development of hardware infrastructure and the availability of massive video datasets, deep learning, an emerging area of machine learning, provides superior cognitive performance in most computer vision tasks (e.g., classification, detection, segmentation, etc.) every year through end-to-end training of deep models. In this subsection, we describe a series of state-of-the-art deep learning methods used to effectively recognize actions in videos. For example, Toudjeu and Tapamo [17] proposed a 2D CNN model to learn spatial features from the temporal information of action videos. However, 2D CNNs cannot model the motion information stored in the input video frames. To this end, the authors of [18], [19] developed the first 3D CNN models to extract and learn the spatial and temporal motion features of multiple adjacent frames. Inspired by the works of [18], [19], several other 3D CNN-based enhanced architectures have been proposed in the last few years to ensure a balance between recognition accuracy and efficiency, including [4], [20]. Since the development of 3D CNN models, they have outperformed 2D CNNs in terms of recognition accuracy for most activity recognition tasks (i.e., HAR tasks) [21]. However, their use has been limited, mainly due to their expensive and time-consuming nature. Recently, only a few attempts have been made to address these limitations by proposing a mixture of 2D and 3D CNNs as a unified network, as in the work of Zhou et al. [22].

III. Proposed Model

The proposed model architecture accomplishes the action recognition task by coordinating three main processing components: a hybrid 2D/3D CNN, an LSTM network, and a visual attention mechanism. Fig. 1 shows each architectural level of the proposed model. First, frame-level discriminative and informative features are extracted by a hybrid 2D/3D CNN as the first pass of recognition. An LSTM network is then used to further model the motion information obtained from the previous representation, helping to capture hidden correlations and long-range dependencies throughout the data sequence. Finally, to improve the contextual representation power of our model, we integrated a top-level attention mechanism that continuously detects the most relevant and salient features of the object of interest. A detailed summary of our proposed architecture is shown in Table I.

TABLE I
OVERVIEW OF THE PROPOSED MODEL WITH ITS LAYERS, OUTPUT SHAPES, AND NUMBER OF PARAMETERS.

Layer	***Output Shape***	***Parameters***
Input Data	(35, 20, 20, 1)	0
Conv3D	(35, 20, 20, 1)	2016
Conv3D	(35, 20, 20, 1)	6928
Conv3D	(35, 20, 20, 1)	6928
Conv3D	(35, 20, 20, 1)	6928
MaxPooling3D	(18, 10, 10, 1)	0
MaxPooling3D	(18, 10, 10, 1)	0
MaxPooling3D	(18, 10, 10, 1)	0
Conv3D	(18, 10, 10, 3)	13856
Conv3D	(18, 10, 10, 3)	13856
Conv3D	(18, 10, 10, 3)	13856
Add	(18, 10, 10, 3)	0
Conv2D	(18, 10, 10, 3)	1056
Conv2D	(18, 10, 10, 3)	1056
MaxPooling2D	(18, 10, 10, 3)	0
Conv2D	(18, 10, 10, 3)	4128
Conv2D	(18, 10, 10, 3)	9248
Conv2D	(18, 10, 10, 3)	1056
Concatenate	(18, 10, 30, 3)	0
Flatten	(18, 9600)	0
Dense	(18, 64)	614464
Dense	(18, 128)	8320
LSTM	(18, 256)	394240
Attention	(128)	65536
Dense	(6)	774
Total	-	1,2M

In the following subsections, each component of the proposed architecture is described in detail.

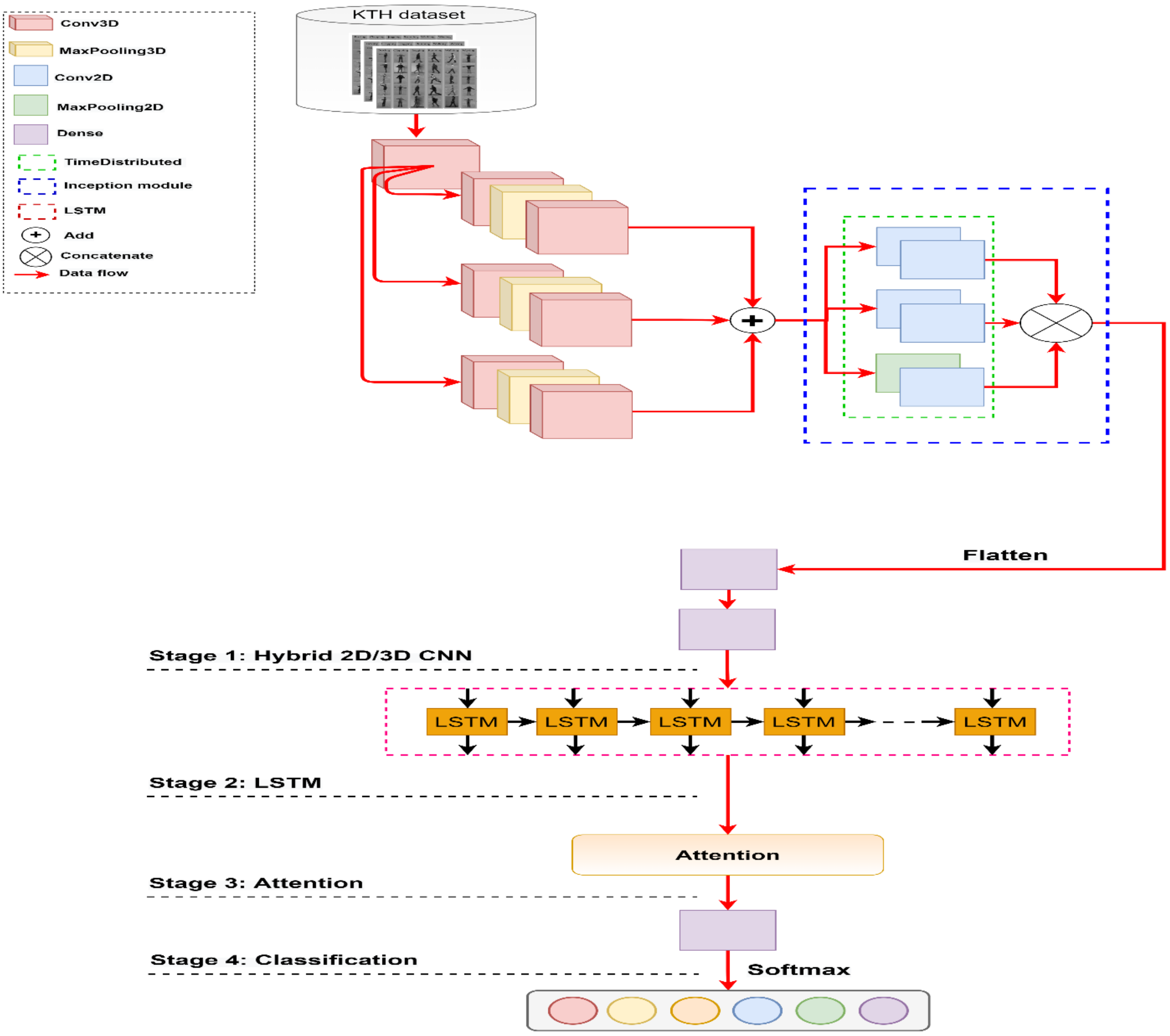


Fig. 1: Overview of the proposed framework for human action recognition using KTH dataset

A. Stage 1: Hybrid 2D/3D CNN for feature extraction

Since its first introduction, CNN has been one of the most widely used discriminative models in computer vision, showing impressive performance in most action recognition systems. These systems typically focus on 2D CNNs due to their computational efficiency and accuracy compared to their 3D counterparts. In contrast, the task of recognizing actions from the video is a temporal classification problem that is highly dependent on the spatial, temporal, and contextual information inherent in the data stream sequence. Currently, there are only a limited number of approaches that use supervised 3D CNN models to classify human actions and activities in video sequences, as their computational complexity makes them unsuitable for most real-time scenarios. To solve this problem, we proposed a hybrid 2D/3D CNN feature extractor that combines 2D and 3D CNNs in a shared structure to model spatio-temporal information and ensure smooth interaction and synergy between multidimensional data spaces. In a 3D CNN, the input data consists of a 3D data cube in which a series of equal-sized kernels can be applied to detect temporal structure and contextual features. Once detected, these feature maps are then fed into an activation function (i.e., non-linearity function) that generates a new set of non-linear activations. By sliding a set of 2D convolutional kernels on the spatial information of the input dimension, highly discriminative 2D feature descriptors can be computed with high computational efficiency. In formal terms, consider a spatio-temporal 3D data volume (e.g., a video sequence) as a multidimensional tensor of the form *[T, H, W, C]*, where T, H, W, and C represent the temporal (contextual) axis, the height and width of the input spatial dimension, and the number of slices, respectively. In 2D CNNs, a set of 2D convolution kernels is applied only to features in the spatial domain of the form *[H, W, C]* to produce a set of feature tensors of the same rank, ignoring the temporal and contextual information contained in multiple consecutive frames. In practical terms, stacking multiple 3D convolutional layers at the bottom of the entire architecture can produce distinct and more complex feature maps, increasing the contextual representation of features.

However, training a deep 3D CNN can affect the speed of the inference process and optimization performance due to excessive memory consumption and computational cost. 2D CNNs can learn spatial features more efficiently than their 3D counterparts, while 3D CNNs are more suitable for modeling motion features and contextual information. Inspired by the above properties, we proposed a hybrid 2D/3D CNN as a robust feature extractor that retains the generalization and discrimination capabilities of 2D and 3D features extracted from spatio-temporal signals. Fig. 1 shows the workflow diagram of our proposed hybrid 2D/3D CNN. We can see that the 3D convolutional side of our hybrid 2D/3D CNN consists of seven 3D convolutional layers (Conv3D) and three max-pooling layers (MaxPooling3D). An activation layer (Rectified Linear Unit (ReLU)) is then added after each convolutional layer to avoid the vanishing gradient problem. Max-pooling layers are used as a dimensionality reduction technique to reduce the spatial dimension of the convolved feature maps. To extract and learn richer contextual patterns and improve the expressiveness of the recognition model, we used a three-level 3D CNN in which the output activations of each level are fused by a merging operator (additive operation). The fused feature maps are then used as input to the 2D convolutional side (2D CNN). Structurally, the 2D CNN consists of a modified Inception module with five 2D convolutional layers (Conv2D), each followed by a ReLU activation function, and only one max-pooling layer (MaxPooling2D), resulting in additional performance improvements. Typically, the number of high-level feature channels generated at each level of abstraction increases sequentially, leading to an expansion in the number of trainable parameters and computational cost. Accordingly, our architecture incorporates an improved TimeDistributed Inception module that performs dimensionality reduction by stacking and combining convolutions with kernels of multiple sizes in each layer for improving computational efficiency. To extract temporal attributes from time-varying data, we used the TimeDistributed wrapper layer by applying the same layer to each time step of many inputs. A flattening layer is then stacked to reshape the input feature space into a fixed-length tensor. Subsequently, to densely connect all the activations of the previous layer, we used two consecutive dense layers followed by batch normalization and dropout layers to ensure the regularization of the training algorithm and improve the generalization ability of the model.

B. Stage 2: LSTM for modelling short- and long-term dependencies

Action recognition in video is a temporal problem that must consider the time-varying nature of the features in each frame. To encode the spatio-temporal characteristics of dynamic signals, we used an LSTM network as an enhanced variant of RNN. Since its first introduction to the community, this variant has shown strong adaptability in modeling both short- and long-range dependencies of sequence features, and avoids the gradient vanishing and exploding problems of ordinary RNNs. Based on its internal memory, an LSTM unit consists of three main elements called gates, namely: a forget gate, an input gate, and an output gate. The role of these gates is to control the flow of information from previous time steps and hidden states. More specifically, when a sequence of input signals is fed to an LSTM network (here, 256 LSTM units), it serves to control the action being performed in its memory cell, either by writing (input gate), reading (output gate), or resetting (forget gate) the state. Fig. 2 illustrates the structure of an LSTM memory cell with three gates.

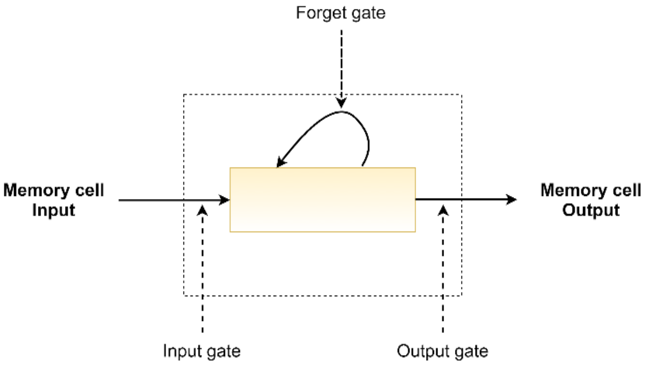


Fig. 2: Schematic representation of an LSTM memory cell

C. Stage 3: Attention mechnaism for detecting salient structures

This work used an attention mechanism that assigns attentional weights to the most relevant regions corresponding to the body parts of moving objects. This strategy would guarantee the discriminability of contextual features and ignore irrelevant patterns. In other words, when encoding a sequence of spatio-temporal cues using the LSTM network, the attention mechanism will capture the most relevant region of each input representation. As shown in Fig. 1, an attention module is integrated into our architecture to select the hierarchical contextual features of the input sequence used for the next prediction step, thus improving the performance of the overall classifier. For this purpose, the attention module first weights the input sequences and then averages them so that the relevant information can be successfully obtained.

D. Stage 4: Action classification

To perform the final classification of the potential actions identified, we added a dense layer with a fixed output size of 6 (i.e., the number of classes to be identified), using Softmax as the activation function (see Fig. 1).

IV. Experiments and Results

In this section, we describe the implementation details of the proposed model, followed by the experimental results obtained using the KTH dataset.

A. Dataset

The performance of the proposed action recognition model was evaluated by conducting experiments using the well-known KTH dataset. The dataset consists of recordings of 25 subjects performing six different human actions (i.e., walking, jogging, running, slapping, waving, and clapping) in four diverse scenarios: outdoors, outdoors with different scales, outdoors with different clothing, and indoors. These scenarios investigated the algorithm's ability to detect human actions, regardless of the actor's background, change in appearance, or body size and shape. The final version of the dataset contains 2391 sequences taken by a still camera at 25 frames per second (fps), each down-sampled with a spatial resolution of 160×120 pixels. In our experiments, we split the video sequences (600 videos in total) into two separate sets: the training/validation set and the testing set. The training/validation set contains 80% of all sequences, while the rest of the samples are included in the testing set. As for the evaluation metrics, we evaluated the prediction performance of the proposed method using the standard accuracy measure.

B. Implementation Setups

This section outlines the implementation details of our model. In this work, we took a 35-frame sequence as input and extracted foreground objects based on hierarchical background subtraction [23]. To reduce memory and hardware requirements, we rescaled the spatial resolution of each input frame from 160×120 to 20×20 pixels. To help maximize the prediction results, we performed several experiments to find the optimal tuning of the hyper-parameters, including the batch size and the loss function. The optimal configuration can be found in Table II.

From Table II, we can see that we used the categorical cross-entropy as a loss function since we are dealing with a multiclass classification problem. In addition, we employed Adam as an efficient optimizer to speed up the optimization process and promote convergence while reducing prediction errors.

For training, we used a laptop with Python 3 and the publicly available Keras library with TensorFlow 2.0 backend. We ran our model on an Nvidia Tesla K80 GPU and about 12 GB of RAM.

TABLE II
OPTIMAL CONFIGURATION OF THE PROPOSED MODEL.

Batch size	15
Number of epochs	100
Loss function	Categorical cross-entropy
Optimizer	Adam

C. Results

To demonstrate the robustness of the proposed hybrid model, we report the classification results in terms of recognition accuracy, as shown in Table III. From Table III, it can be seen that our model achieves an overall recognition accuracy of 96.80% on the KTH dataset, compared to other state-of-the-art methods. Figs. 3 and 4 show the training and validation loss curves and the confusion matrix for the proposed network. In Fig. 3, we can observe that the model converges quickly and stabilizes at a certain level, avoiding the problem of overfitting. As shown in Fig. 4, the vast majority of the testing samples were correctly classified, representing the number of true positives and true negatives in each class. Specifically, our model can classify the actions "*hand waving*", "*jogging*", "*running*", and "*walking*" with 100% accuracy, with some slight confusion between "hand *clapping*" and "*boxing*" (25% and 10%, respectively), mainly due to the similarity between the different action classes. Besides, we can see from Table I that our model was trained with a smaller number of trainable and non-trainable parameters (1,2M parameters in total).

TABLE III
COMPARISON OF CLASSIFICATION ACCURACY WITH STATE-OF-THE-ART METHODS ON THE KTH DATASET.

Papers	*Method*	*Accuracy (%)*
Proposed work	**Hybrid 2D/3D CNN**	**96.80**
[17]	2D CNN	89.17
[4]	3D CNN	94.90
[18]	3D CNN	94.39
[19]	3D CNN	90.20

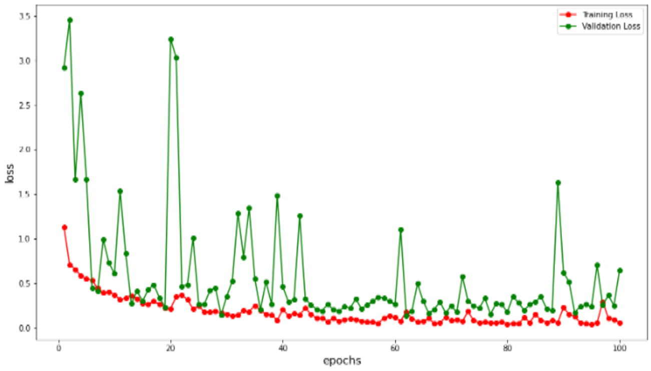
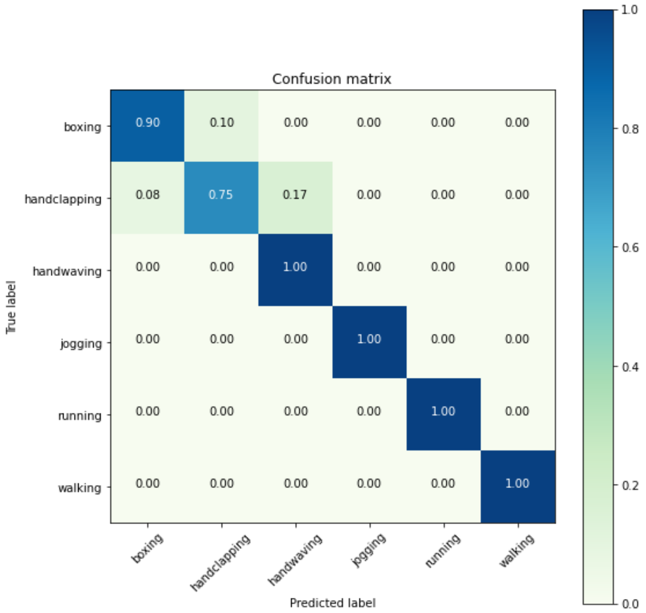


Fig. 3: Illustration of the evolution of training and validation losses per epoch



	boxing	handclapping	handwaving	jogging	running	walking
boxing	0.90	0.10	0.00	0.00	0.00	0.00
handclapping	0.08	0.75	0.17	0.00	0.00	0.00
handwaving	0.00	0.00	1.00	0.00	0.00	0.00
jogging	0.00	0.00	0.00	1.00	0.00	0.00
running	0.00	0.00	0.00	0.00	1.00	0.00
walking	0.00	0.00	0.00	0.00	0.00	1.00

Fig. 4: Confusion matrix of the proposed classifier

V. Conclusion and Future Work

In this paper, we proposed an end-to-end HAR model based on a hybrid 2D/3D CNN. The proposed model combines the hybrid 2D/3D CNN features extracted from the video sequences with an LSTM network to model the short- and long-term dependencies in the data structure. We also used the attention mechanism to improve the contextual representation power of the model. The experimental results performed on the KTH dataset showed that the proposed method outperforms other baseline work in terms of classification accuracy.

In the future, our challenge is to improve the discriminatory performance of the proposed architecture in terms of accuracy and efficiency by better tuning the hyperparameters and enhancing the architectural design of the model.

References

[1] D. R. Beddiar, B. Nini, M. Sabokrou, and A. Hadid, “Vision-based human activity recognition: a survey,” *Multimed Tools Appl*, vol. 79, no. 41, pp. 30509–30555, Nov. 2020, doi: 10.1007/s11042-020-09004-3.

[2] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, Art. no. 7553, May 2015, doi: 10.1038/nature14539.

[3] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, “A survey of the recent architectures of deep convolutional neural networks,” *Artif Intell Rev*, vol. 53, no. 8, pp. 5455–5516, Dec. 2020, doi: 10.1007/s10462-020-09825-6.

[4] J. Arunnehru, G. Chamundeeswari, and S. P. Bharathi, “Human Action Recognition using 3D Convolutional Neural Networks with 3D Motion Cuboids in Surveillance Videos,” *Procedia Computer Science*, vol. 133, pp. 471–477, Jan. 2018, doi: 10.1016/j.procs.2018.07.059.

[5] K. Bayoudh, F. Hamdaoui, and A. Mtibaa, “Hybrid-COVID: a novel hybrid 2D/3D CNN based on cross-domain adaptation approach for COVID-19 screening from chest X-ray images,” *Phys Eng Sci Med*, vol. 43, no. 4, pp. 1415–1431, Dec. 2020, doi: 10.1007/s13246-020-00957-1.

[6] K. Bayoudh, F. Hamdaoui, and A. Mtibaa, “Transfer learning based hybrid 2D-3D CNN for traffic sign recognition and semantic road detection applied in advanced driver assistance systems,” *Appl Intell*, vol. 51, no. 1, pp. 124–142, Jan. 2021, doi: 10.1007/s10489-020-01801-5.

[7] G. Van Houdt, C. Mosquera, and G. Nápoles, “A review on the long short-term memory model,” *Artif Intell Rev*, vol. 53, no. 8, pp. 5929–5955, Dec. 2020, doi: 10.1007/s10462-020-09838-1.

[8] J. Sun, J. Jiang, and Y. Liu, “An Introductory Survey on Attention Mechanisms in Computer Vision Problems,” in *2020 6th International Conference on Big Data and Information Analytics (BigDIA)*, Dec. 2020, pp. 295–300. doi: 10.1109/BigDIA51454.2020.00054.

[9] H. Liu, H. Tang, W. Xiao, Z. Guo, L. Tian, and Y. Gao, “Sequential Bag-of-Words model for human action classification,” *CAAI Transactions on Intelligence Technology*, vol. 1, no. 2, pp. 125–136, Apr. 2016, doi: 10.1016/j.trit.2016.10.001.

[10] C. Li and T. Hua, “Human Action Recognition Based on Template Matching,” *Procedia Engineering*, vol. 15, pp. 2824–2830, Jan. 2011, doi: 10.1016/j.proeng.2011.08.532.

[11] S. H. Kumar and P. Sivaprakash, “New approach for action recognition using motion based features,” in *2013 IEEE Conference on Information Communication Technologies*, Apr. 2013, pp. 1247–1252. doi: 10.1109/CICT.2013.6558292.

[12] H. A. Abdul-Azim and E. E. Hemayed, “Human action recognition using trajectory-based representation,” *Egyptian Informatics Journal*, vol. 16, no. 2, pp. 187–198, Jul. 2015, doi: 10.1016/j.eij.2015.05.002.

[13] L. Fan, Z. Wang, and H. Wang, “Human Activity Recognition Model Based on Decision Tree,” in *2013 International Conference on Advanced Cloud and Big Data*, Dec. 2013, pp. 64–68. doi: 10.1109/CBD.2013.19.

[14] K. G. Manosha Chathuramali and R. Rodrigo, “Faster human activity recognition with SVM,” in *International Conference on Advances in ICT for Emerging Regions (ICTer2012)*, Dec. 2012, pp. 197–203. doi: 10.1109/ICTer.2012.6421415.

[15] X. Yang and Y. L. Tian, “EigenJoints-based action recognition using Naïve-Bayes-Nearest-Neighbor,” in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Jun. 2012, pp. 14–19. doi: 10.1109/CVPRW.2012.6239232.

[16] P. Paul and T. George, “An effective approach for human activity recognition on smartphone,” in *2015 IEEE International Conference on Engineering and Technology (ICETECH)*, Mar. 2015, pp. 1–3. doi: 10.1109/ICETECH.2015.7275024.

[17] I. T. Toudjeu and J.-R. Tapamo, “A 2D Convolutional Neural Network Approach for Human Action Recognition,” in *2019 IEEE AFRICON*, Sep. 2019, pp. 1–5. doi: 10.1109/AFRICON46755.2019.9133840.

[18] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, “Sequential Deep Learning for Human Action Recognition,” in *Human Behavior Understanding*, Nov. 2011, pp. 29–39. doi: 10.1007/978-3-642-25446-8_4.

[19] S. Ji, W. Xu, M. Yang, and K. Yu, “3D Convolutional Neural Networks for Human Action Recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, Jan. 2013, doi: 10.1109/TPAMI.2012.59.

[20] J. Li, Z. Xu, J. Li, and J. Wang, “An Improved Human Action Recognition Method Based on 3D Convolutional Neural Network,” in *Advanced Hybrid Information Processing*, Oct. 2018, pp. 37–46. doi: 10.1007/978-3-030-19086-6_5.

[21] M. Sornam, K. Muthusubash, and V. Vanitha, “A Survey on Image Classification and Activity Recognition using Deep Convolutional Neural Network Architecture,” in *2017 Ninth International Conference on Advanced Computing (ICoAC)*, Dec. 2017, pp. 121–126. doi: 10.1109/ICoAC.2017.8441512.

[22] Y. Zhou, X. Sun, Z.-J. Zha, and W. Zeng, “MiCT: Mixed 3D/2D Convolutional Tube for Human Action Recognition,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun. 2018, pp. 449–458. doi: 10.1109/CVPR.2018.00054.

[23] J. M. McHugh, J. Konrad, V. Saligrama, and P.-M. Jodoin, “Foreground-Adaptive Background Subtraction,” *IEEE Signal Processing Letters*, vol. 16, no. 5, pp. 390–393, May 2009, doi: 10.1109/LSP.2009.2016447.