IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 20, 2018, accepted September 2, 2018, date of publication September 13, 2018, date of current version October 12, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2869790

Human Action Monitoring for Healthcare Based on Deep Learning

YONGBIN GAO[1], XUEHAO XIANG[1], NAIXUE XIONG[2], (Senior Member, IEEE), BO HUANG[1], HYO JONG LEE[3], RAD ALRIFAI[2], XIAOYAN JIANG[1], AND ZHIJUN FANG[1], (Senior Member, IEEE)

[1]School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China
[2]Department of Mathematics and Computer Science, Northestern State University, Tahlequah, OK 74464, USA
[3]Division of Computer Science and Engineering, Center of Advanced Image and Information Technology, Chonbuk National University, Jeonju 54896, South Korea

Corresponding authors: Naixue Xiong (xiongnaixue@gmail.com) and Zhijun Fang (zjfang@sues.edu.cn)

This work was supported in part by the Project of Local Colleges' and Universities' Capacity Construction of Science and Technology Commission in Shanghai under Grant 15590501300, in part by the National Natural Science Foundation of China under Grant 61802253, Grant 61461021, Grant 61603242, and Grant 61772328,in part by the Collaborative Innovation Center for Economic Crime Investigation and Prevention Technology of Jiangxi Province under Grant JXJZXTCX-027 and Grant JXJZXTCX-030, in part by the Shanghai Chenguang Talented Program under Grant 17CG59, in part by the Ministry of Science and ICT (MSIT), South Korea, under the ITRC Support Program supervised by the IITP under Grant IITP-2018-2015-0-00378, and in part by the Basic Science Research Program through the NRF funded by the Ministry of Education under Grant GR2016R1D1A3B03931911.

ABSTRACT Human action monitoring can be advantageous to remotely monitor the status of patients or elderly person for intelligent healthcare. Human action recognition enables efficient and accurate monitoring of human behaviors, which can exhibit multifaceted complexity attributed to disparities in viewpoints, personality, resolution and motion speed of individuals, etc. The spatial-temporal information plays an important role in the human action recognition. In this paper, we proposed a novel deep learning architecture named as recurrent 3D convolutional neural network (R3D) to extract effective and discriminative spatial-temporal features to be used for action recognition, which enables the capturing of long-range temporal information by aggregating the 3D convolutional network entries to serve as an input to the LSTM (Long Short-Term Memory) architecture. The 3D convolutional network and LSTM are two effective methods for extracting the temporal information. The proposed R3D network integrated these two methods by sharing a shared 3D convolutional network in sliding windows on video streaming to capturing short-term spatial-temporal features into the LSTM. The output features of LSTM encapsulate the long-range spatial-temporal information representing high-level abstraction of the human actions. The proposed algorithm is compared to traditional and the-state-of-the-art and deep learning algorithms. The experimental results demonstrated the effectiveness of the proposed system, which can be used as smart monitoring for remote healthcare.

INDEX TERMS Action recognition, 3D convolutional network, LSTM.

I. INTRODUCTION

Since last decade, number of intelligent healthcare applications has dramatically increased due to the problems of aging population and shortage of resources, such as qualified nursing staff and beds in healthcare organizations. Human action recognition enables us to remotely monitor the behavior of patients or elderly person to record their daily activities and ensure safety, such as fall down and abnormal behaviors. The recent advance in artificial intelligence provides the solution for automatically analyze the human action and behaviors. Human action recognition has been investigated for decades due to its high demand in surveillance systems, anomaly detection, video search, etc. Human action recognition is crucial to understand and analyze human activity, in which multiple actions are performed simultaneously. Human action recognition is a challenging task because of the large number of possibilities in intra-class variation that can be attributed to the extensive number of human poses, illuminations [1], occlusion [2], resolution [3], and motion speeds. Thus, an efficient representation of human actions is essential for designing a robust human action recognition system.

An efficient representation of human action needs to be generic, compact, efficient, and simple [4]. The represented features should be general enough to accommodate a wide variety of sources and categories of actions. Due to the huge volume of video data, the video features used in the recognition systems ought to be compact to be suitable for processing, retrieving, and storing. The features also need to be efficient and simple to satisfy the needs of real time applications.

Earlier attempts to develop systems for intelligent analysis of human action relied on hand-crafted features. Informative regions were detected by Space Time Interest Points (STIP) [5], and the temporal Gabor filters were detected by Cuboid [6] algorithm. The traditional local features were extended to include 3D action recognition, such as, SIFT-3D [7] and HOG-3D [8]. These local features are low-level as no semantics and discriminative information were encapsulated. The mid-level and high-level features were extracted by Action Bank [9], Dynamic-Poselets [10], Actons [11], etc. These video representation methods are typically optimized by incorporating heuristic methods to extract discriminative components as features. It is reported that the improved Dense Trajectories (iDT) [12] algorithms achieved the state-of-the-art performance results when compared to other hand-crafted features. Human action is a complex behaviour with large intra-class variation, thus, it needs high-level abstraction of temporal motion information. In this sense, the use of hand-crafted features is inefficient method for extracting high-level abstract information and can fall short in handling the natural variations in human actions. In addition, hand-crafted features are typically high-dimension, and therefore computationally expensive for processing large-scale video data.

Recent advances in computational power(GPU/ clusters) [13], [14] and the availability of large scale training data sets are leading to a growing interest in deep learning research [15]–[17]. Consequently, leading to a breakthrough performance in computer vision tasks including methods for face recognition [18]–[20], speech recognition, natural language processing, etc. ConvNets have been extensively used for both academic research and industry due to its impressive performance. In recent years, ConvNets has also been used for human action recognition, and achieved significant improvement over traditional methods. The ConvNets can be performed in 2D/3D manner. A two-stream ConvNet was used to develop a model for human action recognition where a 2D spatial net was pre-trained on ImageNet, and a temporal net was trained on the optical flow for capturing the motion [21]. Alternatively, a convolutional 3D (C3D) network was proposed to explicitly capture the temporal motion using a 3D convolutional kernel [4]. It is noted that 3D convolution exhibits better performance than 2D methods in providing a straightforward method to extract temporal information. However, the current 3D convolution solution is limited to short clips to accommodate the small size of the 3D kernel.

In this paper, we propose a recurrent 3D convolutional network for human action recognition to capture the long-range temporal structure by aggregating the 3D convolutional network in a LSTM (Long Short-Term Memory) architecture. Our finding suggests that temporal motion is the most informative feature for understanding and analyzing the video frames. It is observed that the 3D convolutional network and LSTM are two effective methods to extract the temporal information. We propose a hybrid network for windowing the frames of a video using 3D convolutional network to obtain short-term spatial-temporal features, and aggregate the features by LSTM to extract the long-range spatial-temporal information conveying a high-level of abstraction of human actions. To our knowledge, this study is the first attempt to aggregate the 3D convolutional network for recurrence analysis of video.

The remainder of this paper is organized as follows. Section 2 describes the related work of human action recognition algorithms, which consists of the traditional and deep learning algorithms. The proposed recurrent 3D convolutional network was introduced in detail in Section 3, which includes the 3D convolutional network, LSTM, and the integration of these two algorithms. Section 4 shows the experimental results on UCF-101 dataset [22], the visualization of features and the comparison results were illustrated in this section.

II. RELATED WORK

Human action recognition has attracted numerous researchers for the past few years, the related researches can be grouped into two categories: traditional methods and convolutional networks.

The vital component of action recognition is capturing the temporal structure of video clips. Gaidon et al. proposed to use ‘‘actoms’’ to semantically characterize the action by a sequence of histograms of actom-anchored features, a nonparametric actom sequence model is then used to detect human action [23]. For complex actions classification, Niebles et al. proposed a latent SVM to differentiate the temporal decomposition of motion segments [24]. The latent temporal decomposition algorithm was then extended based on the latent hierarchical model [25] and the segmental grammar model [26]. Fermando proposed to use a ranking function to learn the evolution of the human appearance over time, and the parameters of this function were regarded as temporal features of the action [27]. Other low-level features were proposed to extend the traditional feature extraction method to 3D. Scovanner et al. proposed a SIFT-3D descriptor to represent video, and bag of words were used to discover spatio-temporal features [7]. Alexander et al. proposed a HOG-3D descriptor for video sequences, which used histograms of oriented 3D spatio-temporal gradients to characterize action [8].

More recently, mid-level and high-level descriptors were proposed to discover the semantic meaning of videos. Action bank was proposed to detect a set of individual action from semantic space and viewpoint space, resulting in highly discriminative performance with simple linear support vector machines [9]. Dynamic-Poselets learn a relational model to decompose human action into temporal "key poses" and spatial "action parts" [10]. Zhu *et al.* [11] proposed to extract a mid-level "actons" feature by multiple instance learning with multi-channel max-margin optimization, the learned actons tends to be more compact, informative, discriminative, and scalable.

Convolutional networks have exhibited impressive performance improvement for action recognition. Karpathy *et al.* [28] evaluated the deep ConvNets on a large-scale dataset Sport1-1M, and the CNNs with two spatial resolutions were extended to video for connectivity in time domain . Simonyan and Zisserman [21] proposed a two-stream ConvNets to separate the spatial from temporal features, the dense optical flow significantly improves the performance despite of limited training data. A trajectory-pooled deep convolutional descriptor (TDD) was proposed to improve the two-stream ConvNets [29], a trajectory-constrained pooling was proposed to combine the learned features. Motion vector based ConvNet was proposed to speed up the two-stream networks by replacing the optical flow with motion vector [30]. Long-range action recognition is also an important subject, Donahue *et al.* [31] proposed a long term recurrent convolutional networks by using the LSTM to extract the long-range temporal information. Tran *et al.* [4] proposed a C3D algorithm to extract spatial-temporal features by 3D convolutional network [34]. Sun *et al.* [35] proposed a factorized spatial-temporal convolutional network with a decomposed 3D convolutional kernels, where the 3D convolutional kernel was factorized as 1D temporal kernel on top of 2D spatial kernels.

Untrimmed videos are common in practical applications, the difficulty of action recognition for untrimmed videos lies on the action localization. A deep segmental network [32] was proposed to split the video into clips, and extract feature for each clip. Wang *et al.* [33] proposed an attention model to recognize the action for untrimmed videos, the model consists of classification module and selection module, the time duration of each action was inferred from the model. Spatial region proposals and temporal action proposals were proposed to localize the action from untrimmed videos. Peng and Schmid [43] proposed a two-stream network to extract high quality spatial proposals. The temporal action proposals were extracted from sampled proposal candidates by a learning sparse dictionaries [44]. Shou *et al.* [45] proposed multi-stage CNNs to improve the accuracy of action temporal localization. Action tube was proposed to spatio-temporal action proposal by multi-box detector [46]. Li *et al.* [47] proposed to regress the action location by an LSTM network based on frame level prediction.

Skeleton data provides a simplified input for action recognition, which describes the high level abstract of human key joints. Skeleton data tends to be more robust to the variations in appearances and location. The prevalence of depth camera, such as Kinect, Realsense, makes the 3D skeleton information accessible. 3D skeleton estimates the 3D coordinates of key joints of the human body. An actionlet model was proposed to extract the features of a subset of key joints [48], the action is then calculated as a linear combination of these actionlet. Orderlet extents actionlet by adding the distance of human joint, and enables the variation in subset size [49]. Song *et al.* [50] proposed to use LSTM to learn the attentions of skeleton joints for different frames.

III. PROPOSED RECURRENT 3D CONVOLUTIONAL NETWORK

In this section, the proposed recurrent 3D convolutional network (R3D) will be described. A sliding window was used in the extraction of the spatial-temporal features of the video clips as shown in Fig. 1. Where a 3D convolutional network with multiple layers is windowing the video. The output of the 3D convolutional network is controlled through the windowing to server an input to the bi-directional LSTM. The outputs of the LSTM forward and backward passes are concatenated and fed into the fully connected layers. The 3D convolutional network extracts the local spatial-temporal features of the video clips, and the LSTM enables us to propagate the previous state into current windows, which enables the network to remember the long-term and short-term states of human action. In addition, the bi-directional LSTM enhances the temporal information by performing twice as many forward and backward passes with supervisory signals.

The proposed network improves the pure 3D convolutional network (C3D) [4] which partitions the video into small clips before completing the analysis by the 3D convolutional network. The C3D approach is not efficient in meeting the requirements of practical applications as the long-range temporal information can't be extracted. The proposed network also differs from the LSTM algorithm by Donahue *et al.* [31] or recurrent neural network [51] which doesn't take the advantage of 3D convolutional network or the sliding window structure.

A. 3D CONVOLUTIONAL NETWORK

3D convolutional network is an extension of 2D convolutional network, in which the kernel is three dimensional as shown in Fig. 2. The output of 3D convolutional network is a three-dimensional model, while the output of 2D convolutional network is a 2D data matrix. It is noted that we can concatenate the 2D frame to feed into the 2D ConvNet to capture the temporal motion, while the temporal information disappear during the next 2D ConvNet. In contrast, the 3D convolutional network will explicitly maintain the temporal information since each output feature map has three dimensions consisting of both spatial and related temporal information. Thus, we claim that the 3D convolutional network is

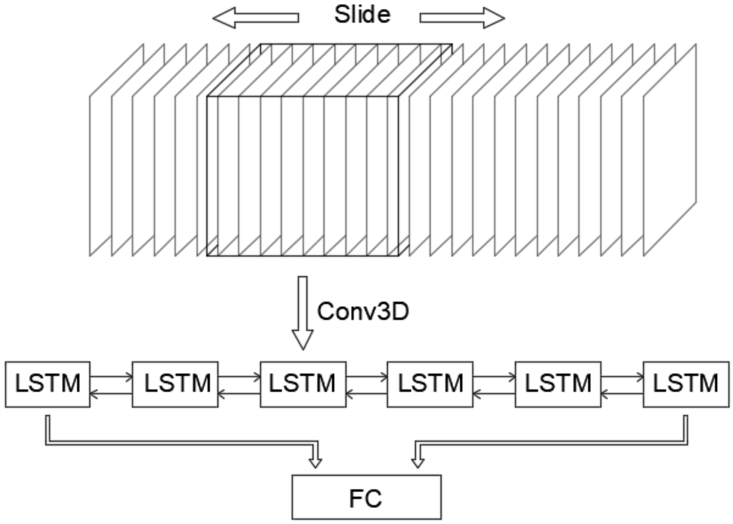


FIGURE 1. Proposed recurrent 3D convolutional network.

TABLE 1. Parameter details of 3D convolutional network.

type	patch size/stride	output size	params	ops
3d convolution	3×3×3/1	32×32×10×32	2.6K	19M
3d convolution	3×3×3/1	32×32×10×32	28K	202M
max pooling	3×3×3/3	11×11×4x32		
3d convolution	3×3×3/1	11×11×4x64	55K	9M
3d convolution	3×3×3/1	11×11×4x64	111K	17M
max pooling	3×3×3/3	4×4×2×64		
fully connected		512	1048K	1M

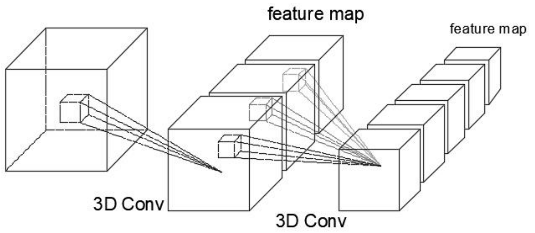


FIGURE 2. Illustration of 3D convolutional network with 3D kernel.

capable of preserving the temporal information while extracting the spatial features. 3D convolutional network is a more straight-forward method to extract spatial-temporal features than the two-stream network, in which two networks were used to extract both the spatial and temporal features from optical flow.

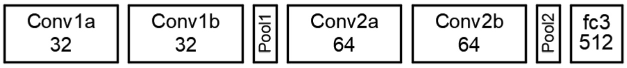


FIGURE 3. A 3D convolutional network for each sliding window.

Multiple 3D convolutional layers were stacked to generate an efficient 3D convolutional network as shown in Fig. 3. The number of 3D convolutional layers is a key parameter for 3D convolutional network. The VGGNet typically achieved favorable performance, while the complexity of VGGNet is high. Thus, the direct extension of VGGNet by altering the 2D convolutional layer with 3D convolutional layer is intractable, especially for the additional LSTM architecture. To enable real time application of the 3D ConvNets, we use a simplified VGGNet, which consists of twice “CONV+CONV+POOLING” followed by a fully connected layer. The first two 3D ConvNets have 32 feature maps, and the next two 3D ConvNets use 64 feature maps. Spatial features also contain context information that can provide relevant clues for improving action recognition. The output of this simplified network is a set of spatial-temporal features represented as a 512-dimensional vector. Then, the output of the previous step is inserted into the LSTM network to learn long-range features. Based our experimental results, our network is much simpler to use yet as efficient as C3D [4].

Suppose that the input data of layer l is X^l, the 3D kernel of each filter bank is $K \in \mathbb{R}^{K_1,K_2,K_3}$, the feature map of 3D convolutional layer $l+1$ is represented as:

$$(X * K)_{ijz} = \sum_{m=0}^{K_1-1} \sum_{n=0}^{K_2-1} \sum_{l=0}^{K_3-1} K_{m,n,l} X_{i+m,j+n,z+l} \tag{1}$$

We use $3 \times 3 \times 3$ as the kernel size of 3D convolutional layer, the detailed network parameters are shown in Table 1.

The proposed network is significant smaller than C3D to enable a real time deployment, yet the recognition accuracy is much higher than C3D due to the additional LSTM mechanism.

The size of sliding window is an important factor for action recognition, a large window size gains the temporal information while significant increase the computational complexity. We empirically set the sliding window in our experiments to k = 10. Let X_t be the t-th clip with k frames of the input video sample, the output of the convolutional network is Y_t.

$$Y_t = C3D\,(X_t) \tag{2}$$

where $C3D$ is the mapping function of 3D convolutional network, which can be calculated by the deep learning model. Each sliding window corresponds to a short clip of 10 frames, and the output turns to be 512 feature values, thus, the C3D model is able to extract high-level abstract of spatial-temporal features of human action.

B. RECURRENT 3D CONVOLUTIONAL NETWORK

Long short term memory is an efficient recurrent neural network, which integrates the long term and short term states of the current processing. Due to its impressive performance, LSTM is extensively used in speech recognition, natural language processing, image captioning. For action recognition, the 3D convolutional network is used to extract local spatial-temporal features, as human action is usually long-range. To extract long-range spatial-temporal features, we aggregate the output of 3D convolutional network in a LSTM architecture as shown in Fig. 4, where one unit of the LSTM has been illustrated, the output of the 3D convolutional network Y_t is fed into the LSTM unit. The LSTM consists of three steps: forget part of the previous state, update the memory cell, and output the state.

The memory cell needs to forget some part of previous state by a forget gate f_t:

$$f_t = \sigma(W_f\,[h_{t-1}, Y_t] + b_f) \tag{3}$$

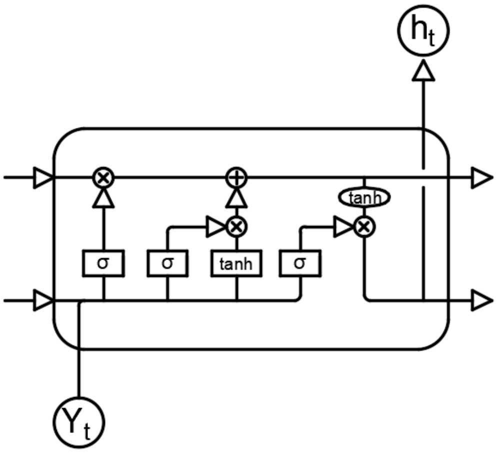


FIGURE 4. Architecture of a LSTM unit.

where f_t is a number between [0, 1] determining the forget percentage of the previous states. The next step is to update the memory cell, which is calculated by the i_t, $\hat{C}_t$:

$$i_t = \sigma(W_i\,[h_{t-1}, Y_t] + b_i) \tag{4}$$

$$\hat{C}_t = \tanh(W_C\,[h_{t-1}, Y_t] + b_C) \tag{5}$$

The i_t, $\hat{C}_t$ is then updated into the memory cell as:

$$C_t = f_t * C_{t-1} + i_t * \hat{C}_t \tag{6}$$

The last step is to determine the output state h_t by the output gate, which is performed as:

$$o_t = \sigma\,(W_o\,[h_{t-1}, Y_t] + b_o) \tag{7}$$

$$h_t = o_t * \tanh(C_t) \tag{8}$$

After the LSTM propagates all the states of spatial-temporal features in a sliding window, it outputs the final state to a fully connected layer. By using the bi-direction LSTM, the outputs of forward and backward pass were concatenated to enhance the temporal information.

Lastly, the softmax loss was used as the supervisory information for training the end-to-end network. The softmax loss was added on top of concatenated features $l \in \mathbb{R}^{512\times2}$ of bi-direction LSTM layer:

$$L = -\sum\nolimits_{j=1}^{K_c} y_j log \frac{e_{l_j}}{\sum_{k=1}^{1024} e_{l_k}} \tag{9}$$

Once the above loss function is defined, the model can be trained based on back-propagation algorithm. The gradient of model weights W can be calculated by a back-propagation chain via $\partial L/\partial W$, the classification errors are back-propagated from the top layers to the bottom layers, and the weights can be updated until the model is converged.

Moreover, the extracted spatial-temporal features from the LSTM were fed into the linear SVM as a classifier. The SVM aims to maximum the margin between two classes:

$$\max_{w,b} \frac{2}{\|w\|} \quad s.t.\, y_i(w^T x_i + b) \geq 1 \tag{10}$$

SVM are inherently bi-class classifier, which requires to build C one-versus-rest classifiers to enable C-class classification problem. Alternatively, linear SVM recognizes the C-class in only one classifier, which significantly reduces the recognition complexity.

IV. RESULTS

We evaluated our proposed R3D network on a large-scale action datasets UCF101, which is the most widely used datasets for human action recognition. The UCF101 contains 13,320 video clips for 101 action classes. The dataset was divided equally into training and validation sets. To enable real time action recognition, we resize the frame size to 32 × 32. The proposed network was implemented on the tensorflow [36] with CUDA support, the batch size was set to 16, the maximum epoch was set to 100, and the network weights were initialized by values driven from Gaussian distribution.

The Adam optimizer was used for training the network [37]. The experiments were conducted on the Nvidia GTX 1080 GPU.

The network parameters were first chosen empirically by examining the performance for various parameters setting. The action recognition performances for the varying kernel size, pooling size, and sliding window size were evaluated. We achieved the best performance when the kernel size was 3×3×3 and the pooling size was 3×3×3. For the sliding window, the best size we used is usually a set of 8-10 frames and a sliding window of size 10 achieved the best performance from our experimental results. TABLE 2 shows the action recognition performance with various sizes of sliding window.

TABLE 2. Action recognition performance by varying sliding window size.

No. Frame	8	9	10
Performance	82.0	85.1	85.7

Different from how the two-stream network is trained, our network is trained from scratch using various network architectures including Alexnet [38], VGGNet [39], GoogleNet [40], and our proposed simplified VGGNet, which exhibits the best performance while the other networks more or less suffer from the overfitting problem. Fig. 5 and Fig. 6 dedicate the loss and accuracy plot for the training and validation process. Our simplified VGGNet is much smaller compared with the primitive VGGNet, it only has twice "CONV+CONV+POOLING" combined layers, while primitive VGGNet involves five "twice "CONV+CONV+POOLING" concatenated. When VGGNet was applied to video, it tends to be overfitting due to the temporal redundancy in video.

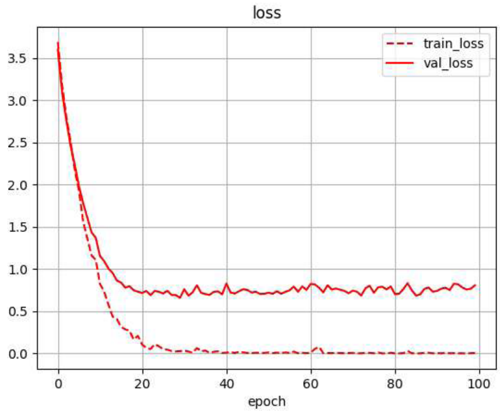


FIGURE 5. Training and validation loss on the UCF-101 dataset.

To understand what the network learned for the action recognition task, we visualized the features map using deconvoluation network as proposed by Zeiler and Fergus [41]. The maximum activated feature map over the samples were back-transformed to the pixel space as illustrated in Fig. 7.

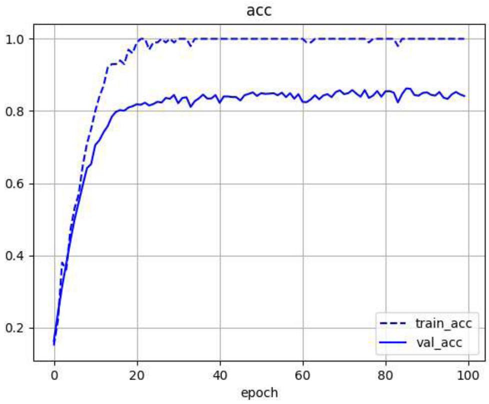


FIGURE 6. Training and validation accuracy on the UCF-101 dataset.

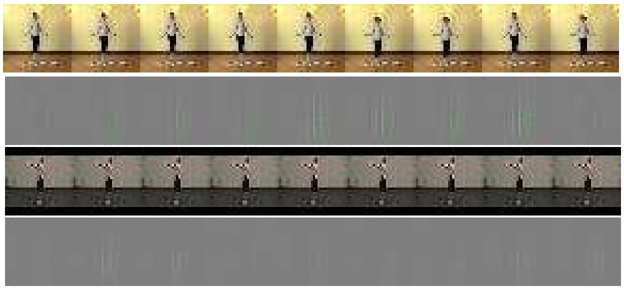


FIGURE 7. Visualization of feature map using deconvolutional network.

We found that the network was able to recognize significant motions since the activated feature map was trained to identify the responses of large motions across frames. In Fig. 7, the first video clip is "jump rope" which shares a similar pattern at some point with the second clip "body weight squats." The Figure illustrated the deconvolutional results of the same feature map that activated for that pattern. while the motion in "jump rope" is stronger than in "body weight squats" clip. As a result, the response is more significant for "jump rope" clip.

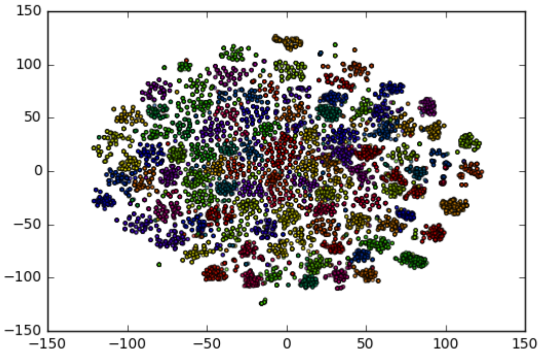


FIGURE 8. Visualization of learned spatial-temporal features.

The extracted spatial-temporal features are visualized using t-SNE tools [42], the high-dimensional features were projected into 2D space as shown in Fig.8. In this figure, the same color corresponds to the same action class, the strong grouping capability of the same class was exhibited in this figure, indicating that the features learned from spatial-temporal input were discriminative.

Lastly, to evaluate our action recognition R3D network, we compare its performance with those reported by other studies as shown in Table 3. The compared algorithms include traditional (iDT [12]), and state-of-the-art including the latest deep learning based methods for action recognition: Deep networks [28], LRCN [31], LSTM composite model [43], C3D [4]. Among these algorithms, LRCN and LSTM composite model are LSTM based algorithms, C3D is a 3D convolutional network. The proposed algorithm achieved the best performance among all the compared algorithms. The experimental results demonstrate the effectiveness of the proposed R3D network. Specially, compared with C3D method that consists of 8 convolutional layers and 5 pooling layers, our proposed R3D only has 4 convolutional layers and 2 pooling layers, our network is much simpler while more accurate than C3D due to the additional LSTM structure. For the traditional C3D algorithm, the computational time is usually high due to the dense trajectories calculation. The running times (frames per second) of compared algorithms are presented in TABLE 3.

TABLE 3. Performance comparison with state-of-the-art algorithms.

Methods	Accuracy	Time
iDT[12]	76.2	4fps
Deep networks [28]	65.4	58fps
LRCN[31]	82.9	298fps
LSTM composite model [43]	84.3	312fps
C3D[4]	82.3	313fps
R3D (Ours)	85.7	427fps
R3D+Linear SVM (Ours)	86.8	386fps

V. CONCLUSION

In this paper, we proposed a recurrent 3D convolutional network to remotely monitor the human action for intelligent healthcare, which is the first attempt to integrate the 3D convolutional network in a recurrent manner. The integrated network uses a shared 3D convolutional network to slide in video to obtain short-term spatial-temporal features, and aggregate the features by LSTM to extract the long-range spatial-temporal information, which represents a high-level abstraction of human actions. We use the deconvolutional network to analyze the learned feature map, and use the t-SNE tools to visualize the learned spatial-temporal features. The results showed that the extracted spatial-temporal features captured the motion in the video, while maintained its discriminative capability needed to recognize the action. The proposed network was compared with traditional methods and deep learning based methods for action recognition, which shows a significant improvement over traditional methods. Also, the proposed network exhibits higher performance than both the 3D convolutional network and LSTM based network. In general, the proposed R3D network is an effective and discriminative algorithm for human action recognition, which can be applied to remotely monitor the patients or elderly person for intelligent healthcare.

REFERENCES

[1] Z. Xia, X. Wang, X. Sun, Q. Liu, and N. Xiong, "Steganalysis of LSB matching using differences between nonadjacent pixels," *Multimedia Tools Appl.*, vol. 75, no. 4, pp. 1947–1962, Feb. 2016.

[2] Y. Fang, Z. Fang, F. Yuan, Y. Yang, S. Yang, and N. N. Xiong, "Optimized multioperator image retargeting based on perceptual similarity measure," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 47, no. 11, pp. 2956–2966, Nov. 2016.

[3] N. Xiong, R. W. Liu, M. Liang, D. Wu, Z. Liu, and H. Wu, "Effective alternating direction optimization methods for sparsity-constrained blind image deblurring," *Sensors*, vol. 17, no. 1, p. 174, 2017.

[4] 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.*, Dec. 2015, pp. 4489–4497.

[5] I. Laptev, "On space-time interest points," *Int. J. Comput. Vis.*, vol. 64, nos. 2–3, pp. 107–123, 2005.

[6] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. 2nd Joint IEEE Int. Workshop IEEE Vis. Surveill. Perform. Eval. Tracking*, Oct. 2005, pp. 65–72.

[7] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. 15th ACM Int. Conf. Multimedia*, 2007, pp. 357–360.

[8] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. 19th Brit. Mach. Vis. Conf. Assoc. (BMVC)*, 2008, pp. 275-1–275-10.

[9] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1234–1241.

[10] L. Wang, Y. Qiao, and X. Tang, "Video action detection with relational dynamic-poselets," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 565–580.

[11] J. Zhu, B. Wang, X. Yang, W. Zhang, and Z. Tu, "Action recognition with actons," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3559–3566.

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

[13] X. Lu, L. Tu, X. Zhou, N. Xiong, and L. Sun, "ViMediaNet: An emulation system for interactive multimedia based telepresence services," *J. Supercomput.*, vol. 73, no. 8, pp. 3562–3578, 2016.

[14] N. Xiong *et al.*, "A novel self-tuning feedback controller for active queue management supporting TCP flows," *Inf. Sci.*, vol. 180, no. 11, pp. 2249–2263, 2010.

[15] Y. Gao and H. J. Lee, "Local tiled deep networks for recognition of vehicle make and model," *Sensors*, vol. 16, no. 2, p. 226, 2016.

[16] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 1701–1708.

[17] Y. Sun, D. Liang, X. Wang, and X. Tang. (2015). "DeepID3: Face recognition with very deep neural networks." [Online]. Available: https://arxiv.org/abs/1502.00873

[18] Y. Gao and H. J. Lee, "Learning warps based similarity for pose-unconstrained face recognition," *Multimedia Tools Appl.*, vol. 77, no. 2, pp. 1927–1942, 2017.

[19] Y. Gao and H. J. Lee, "Cross-pose face recognition based on multiple virtual views and alignment error," *Pattern Recognit. Lett.*, vol. 65, pp. 170–176, Nov. 2015.

[20] Y. Gao and H. J. Lee, "Pose-invariant features and personalized correspondence learning for face recognition," *Neural Comput. Appl.*, pp. 1–10, 2017. [Online]. Available: https://doi.org/10.1007/s00521-017-3035-3

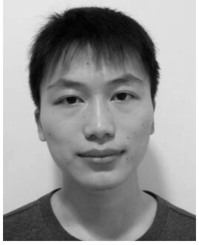
[21] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

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

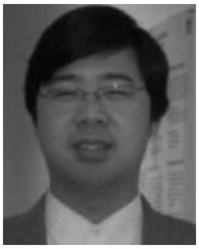
[23] A. Gaidon, Z. Harchaoui, and C. Schmid, "Temporal localization of actions with actoms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2782–2795, Nov. 2013.

[24] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2010, pp. 392–405.
[25] L. Wang, Y. Qiao, and X. Tang, "Latent hierarchical model of temporal structure for complex activity classification," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 810–822, Feb. 2014.
[26] H. Pirsiavash and D. Ramanan, "Parsing videos of actions with segmental grammars," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2014, pp. 612–619.
[27] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5378–5387.
[28] 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.
[29] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 4305–4314.
[30] B. Zhang, L. Wang, Z. Wang, Y. Qiao, and H. Wang, "Real-time action recognition with enhanced motion vector CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2718–2726.
[31] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2625–2634.
[32] L. Wang *et al.*, "Temporal segment networks: Towards good practices for deep action recognition," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 20–36.
[33] L. Wang, Y. Xiong, D. Lin, and L. Van Gool. (2017). "Untrimmednets for weakly supervised action recognition and detection." [Online]. Available: https://arxiv.org/abs/1703.03329
[34] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.
[35] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4597–4605.
[36] M. Abadi *et al.* (2016). "TensorFlow: Large-scale machine learning on heterogeneous distributed systems." [Online]. Available: https://arxiv.org/abs/1603.04467
[37] D. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: https://arxiv.org/abs/1412.6980
[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[39] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[40] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1–9.
[41] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 818–833.
[42] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[43] X. Peng and C. Schmid, "Multi-region two-stream R-CNN for action detection," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 744–759.
[44] F. CabaHeilbron, J. C. Niebles, and B. Ghanem, "Fast temporal activity proposals for efficient detection of human actions in untrimmed videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1914–1923.
[45] Z. Shou, D. Wang, and S. Chang, "Temporal action localization in untrimmed videos via multi-stage CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1049–1058.
[46] G. Singh, S. Saha, M. Sapienza, P. Torr, and F. Cuzzolin, "Online real-time multiple spatiotemporal action localisation and prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3657–3666.
[47] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, and J. Liu, "Online human action detection using joint classification-regression recurrent neural networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 203–220.
[48] J. Wang, Z. Liu, 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.
[49] R. Vemulapalli, F. Arrate, and R. Chellappa, "R3DG features: Relative 3D geometry-based skeletal representations for human action recognition," *Comput. Vis. Image Understand.*, vol. 152, pp. 155–166, Nov. 2016.
[50] S. Song, C. Lan, J. Xing, W. Zeng, and J. Liu, "Spatio-temporal attention-based LSTM networks for 3D action recognition and detection," *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3459–3471, Jul. 2018.
[51] N. Srivastava, E. Mansimov, and R. Salakhudinov, "Unsupervised learning of video representations using LSTMs," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 843–852.

YONGBIN GAO received the Ph.D. degree from Chonbuk National University, South Korea. He is currently a Faculty Member of the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, China. He is selected as a Chenguang Talented Scholar of Shanghai. He has published numerous SCI papers in prestigious journals, such as *Information Science*, *Pattern Recognition Letters*, in the field of image processing, patter recognition, and computer vision.



XUEHAO XIANG is currently pursuing the bachelor's degree under the supervision of Prof. Z. Fang. He was with the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, China. His research area is image processing, pattern recognition, and computer vision.



NAIXUE XIONG (SM'12) received the Ph.D. degrees in software engineering from Wuhan University and in dependable networks from the Japan Advanced Institute of Science and Technology. Before he attends SWOSU, he was a Full Professor with Colorado Technical University for four years, an Assistant Professor with Wentworth Technology Institution for one year, and an Assistant Professor and held a post-doctoral position with Georgia State University for four years. He is currently with the Department of Mathematics and Computer Science, Northestern State University, Tahlequah, OK, USA.

He published over 100 international journal papers and over 100 international conference papers. Some of his works were published in IEEE JSAC, IEEE or ACM Transactions, ACM Sigcomm Workshop, IEEE INFOCOM, ICDCS, and IPDPS. His research interests include cloud computing, business networks, security and dependability, parallel and distributed computing, and optimization theory. He has been a PC member, an OC member, a General Chair, a Program Chair, and a Publicity Chair of over 100 international conferences. He is a Senior Member of the IEEE Computer Society. He has received the Best Paper Award in the 10th IEEE International Conference on High Performance Computing and Communications 2008 and the Best student Paper Award in the 28th North American Fuzzy Information Processing Society Annual Conference 2009. He is the Chair of Trusted Cloud Computing Task Force, the IEEE Computational Intelligence Society, and the Industry System Applications Technical Committee. He was serving as an Editor-in-Chief, an Associate Editor, or an Editor member for over 10 international journals (including an Associate Editor for the IEEE TRANSACTIONS ON SYSTEMS, MAN AND CYBERNETICS: SYSTEMS,an Associate Editor for *Information Science*, an Editor-in-Chief for the *Journal of Internet Technology*, and an Editor-in-Chief for the *Journal of Parallel and Cloud Computing*), and a guest editor for over 10 international journals, including the IEEE WIRELESS COMMUNICATIONS, the IEEE SENSOR JOURNAL, WINET, and MONET.

BO HUANG received the Ph.D. degree from Wuhan University, China. He is currently a Faculty Member of the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, China. His research interests include software engineering, modeling and inference, and image processing. His group received the first prize of Science and Technology Progress Award, Wuhan.

HYO JONG LEE received the B.S., M.S., and Ph.D. degrees in computer science from the University of Utah, USA, in 1986, 1988, and 1991, respectively. Since 1991, he has been with the School of Computer Science and Engineering, Chonbuk National University, as a Professor. He published over 100 journal and conference papers. His current research interests include image processing, computer vision, medical imaging, and parallel processing.

RAD ALRIFAI is currently an Associate Professor of the Department of Mathematics and Computer Science, Northestern State University, USA. His current research interests include image processing and computer vision.

XIAOYAN JIANG received the Ph.D. degree in computer science from the Friedrich-Schiller University of Jena, Jena, Germany. She is currently a Lecturer with the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai, China.

She has published numerous SCI/EI papers in the field of computer vision. Her research interests include multi-object tracking for autonomous driving and visual surveillance, probability theory, and optimization algorithms. She received the fund from the National Natural Science Foundation of China in 2017. She got scholarships from both Chinese Government and German Academic Exchange Service (DAAD).

ZHIJUN FANG (SM'17) received the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China. He is currently a Professor and the Dean of the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai. He has published over 70 papers on prestigious journals and conferences, which includes the highly recognized journals and conferences, such as *Information Sciences*, the IEEE Transsctions on Systems, Man, and Cybernetics, and the IEEE Transactions on Image Processing. He serves as the member/executive directors of many professional societies involving the multimedia, management, education, and artificial intelligence. He is a senior member the China Computer Federation. He is the committee member of many conferences, and specially, he is the chair/co-chair of over 10 conferences including ICMeCG, ISITC, and HHME.