A REVIEW ON HUMAN ACTION RECOGNITION

[1]Archana N., [2]Hareesh K.
[1]M. Tech, Signal Processing and Embedded Systems, Government College of Engineering, Kannur
archanavazhayil@gmail.com
[2]Assistant Professor,Department of Electronics and Communication Engineering
Government College of Engineering, Kannur, hareesh@gcek.ac.in

Abstract—Human action recognition is a growing technology in computer vision. At the same time, it makes so many challenges itself. The applications like surveillance footage, user interfaces, automatic video organization shows the significance of human action recognition. There is a wide range of approaches to recognize human action. With the advancement of dense sampling, dense trajectories are used for the representation of videos. Using Convolutional Neural Networks (CNNs) and Long short-term memory Units (LSTM) we can extract the Spatiotemporal features to identify human action. Temporal Segment Connection Network (TSCN) is used to ensure the continuity of action information between action samples. To avoid large memory consumption and privacy problems Wi-Fi-based action recognition techniques are used.

Index Terms—Human action recognition, Dense Trajectories, Spatio-Temporal Features, Convolution neural network (CNN), Temporal Segment Connection Network (TSCN), Device-free wireless sensing

I. INTRODUCTION

When considering computer vision, human action recognition is a challenging task. The input for Human action recognition is videos, images, and others, like skeleton data, sensor data, etc. The data sources used for human action recognition are Controlled collection data like the Weizmann dataset, KTH dataset etc. And the wild sense data from movies (Eg.HMDB51), YouTube videos (Eg.UCF 101, ActivityNet, Kinetic), and surveillance footage (Eg.HiEve) can also be used. Accuracy is one of the main challenges in human action recognition. Action recognition from a compressed video is still a standard problem. After all, there is an uncertainty in human action recognition model implementation on a new domain.

The present paper has been organized in the following way. Section II outlines human action recognition. The literature review is outlined in Section III. Applications are considered in Section IV. At last, the paper is concluded by Section V.

II. HUMAN ACTION RECOGNITION

In recent years, many kinds of action representation methods have been proposed, including local and global features based on temporal and spatial changes, trajectory features based on key point tracking , motion changes based on depth information , and action features based on human pose changes. Today, there are so many methods of recognizing human action such as the method based on spatial temporal interesting points (STIP), approaches to the analysis of human walking, methods based on deep learning as well as the latest approach called depth learning methods.

RGB data are the main data type used in many human action recognition methods. At the same time skeleton data and depth data are also used. Invention of depth camera could enhance the usage of depth data in human action recognition. Some of the recent works are supported by the fusion data type usage also. Another important aspect of research on the recognition of human actions is that most studies have focused on representations of human action features. Interaction recognition and action detection are the main key problems in human action classification. Interaction is defined as actions involving more than two persons or actions between persons and objects. Action detection is the localization of the position where an action occurs over time and space from image sequence data that has not been segmented.

III. REVIEW OF LITERATURE

Various methods are being proposed for carrying out human action recognition. In the upcoming sections, some of the significant methods are reviewed.

A. Dense Trajectories for Action Recognition

Heng Wang et al.[1] introduced a methodology used to recognize the action by use of dense trajectories. Dence sampling is the canonic advanced technology for this method of action recognition. Human action recognition sometimes shows difficulty during the concern of fast irregular action. This problem is addressed here. Dense trajectories are used to represent the motion information of the input video.The descriptor used here is related to the motion boundary histogram (MBH).

As shown in Fig. 1[1], dense trajectories are mined from multiple spatial scales. A dense sampling of the sampling step size of 5 is used here and observed better results. By using median filtering in a dense optical flow field $f = (r_t,s_t)$, each point of a frame at t that is, $Pt = (x_t,y_t)$ can catch the next frame at t+1 .

$$P(t+1) = (x(t+1),y(t+1)) = (xt,yt)+(N * f)|(\bar{}xt,y\bar{}t) \quad (1)$$

N is the median filtering kernel. Then the trajectories are mapped as $(P_t,P_{t+1},P_{t+2}.....)$.

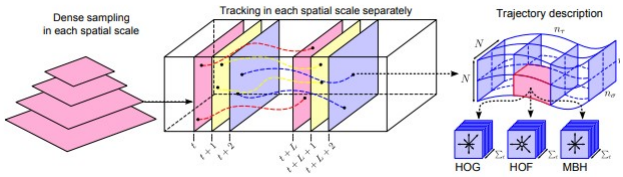

Fig. 1. Representation of the dense trajectory description. [1]

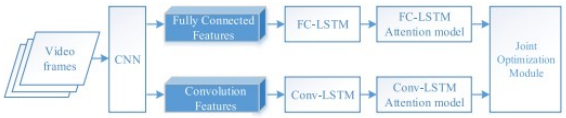

Fig. 2. Flow graph of Spatio-temporal feature extraction [2]

The reuse of dense optical flow fields could decrease the computational complexity. A typical issue in the following is floating. Directions will be in a general float from their underlying area during the tracking. To maintain a strategic distance from this issue, it limits the length of a trajectory to L outlines. When a direction surpasses length L, it is taken out from the tracking cycle, see Figure 1[1] (center). To guarantee a dense inclusion of the video, it checks the presence of a track on the dense lattice in each edge.

The shape of a trajectory encodes local motion patterns. Given a trajectory of length L, and describe its shape by a sequence $S = (\Delta P_t,....\Delta P_{t+1-L})$ of displacement vectors $\Delta P_t = (P_{t+1} - P_t) = (x_{t+1} - x_t, y_{t+1} - y_t)$. The resulting vector is normalized by the sum of the magnitudes of the displacement vector:

$$\vec{S} = \frac{(\Delta P_t, ..., \Delta P_{t+L-1})}{\sum_{j=t}^{t+L-1} \|\Delta P_j\|} \quad (2)$$

This vector is derived from the trajectory descriptor. Representing trajectories are discussed at multiple temporal scales, for recognizing actions with different speeds. After all, this did not enhance the results in practice. Therefore, trajectories with a fixed length L are used.

Optical flow computes the absolute motion, which inevitably includes camera motion[6]. Dalal et al.[7] proposed the MBH (motion boundary histogram) descriptor for human detection, where the derivatives are computed separately for the horizontal and vertical components of the optical flow. This descriptor encodes the relative motion between pixels. In this method, MBH is used to describe dense trajectories [12].

B. Human Action Recognition by Learning Spatio-TemporalFeatures With Deep Neural Networks

It is a deep neural network-based method [9]. RGB data is used as the input data here. The computational complexity of neural networks is very high while considering the optical data. But here the RGB data and its features could decrease this complexity in computation. It also has a thin architecture that consists of different layers of convolution neural network (CNN), Long short-term memory (LSTM) units, Attention models, joint optimization module (JOM) that consist of classifiers like softmax layer, dropout mechanism [2], etc as shown in Fig 2[2]. The advancement of deep learning techniques could satisfy the required accuracy for applicationspecific human activity identification and labeling. Resnet, VGG16, GoogleNet are the main networks used to implement these kinds of action recognition. Here the spatial features are extracted by the convolutional layers of CNN. LSTM units are used to extract the temporal features of the RGB data. The video sequence samples are transformed into frames and these frames are provided to the CNN. The CNN consists of

mainly three layers termed as a convolutional layer, ReLU layer, and a pooling layer. The convolutional layer and the fully connected layer of CNN extract the different features of the input data and that are used to represent the video. The outputs from these two layers provide two distinct feature maps. These two feature maps are given to two separate LSTM-Attention model.

The architecture consists of two types of LSTM (fully connected LSTM and convolutional LSTM) units and also two types of LSTM-Attention models (Fully connected-LSTM Attention model and convolutional LSTM-Attention model). LSTM units can remember the past outputs and also it could predict the output based on that memory. The LSTM unit considers particular states termed as cell states and hidden states. The output of this LSTM unit is provided by the condition of the hidden state. The hidden state depends on the previous state and also the cell state. The cell state provides the output as the function of inputs of the LSTM and by the memory information.

The feature maps are given to the LSTM units to extract the temporal features. The output gate of the LSTM unit is given to the fully connected layer in the fully connected layer LSTM attention model. The output of the fully connected layer is a K dimensional vector if there is, K number of filters in the convolution layer. Then an activation function like tanh could produce a mid result of this vector. This mid result value is given to the softmaxlayer of JOM. It could be identify the probability of each action category. However, this is not sufficient to analyze spatial information. Then the other LSTM-attention model, the convolutional LSTM-Attention model [8] is used to address this information analysis. Here convolutional operations are used in the place of the state to state transition of the previous one. The attention models help to figure out the significant frames required to achieve accurate action recognition

c. LSTM-CNN Architecture for Human Activity Recognition

This architecture provides an encapsulation between the recurrent neural networks, long short term memory with convolutional neural network CNN. The pooling layers like max-pooling layer and normal average pooling layers reduce the dimensionality of the feature maps. A special type of pooling layer called the global average pooling layer (GAP) is used, here to minimize the model parameters. Another special layer called Batch Normalization (BN) layer is used to increase the training speed of the network in the training period. The architecture consists of two LSTM layers, two convolutional layers, and a global average pooling layer, and a batch normalization layer as shown in Fig 3 [3]. After the batch normalization layer, a combination of a fully connected layer and softmax layer provide the output.

Here data from the motion sensors [13] that are placed on the body parts are used as the raw data. And these data are collected by advanced wireless technology. Before providing this data to the network, some preprocessing procedures are done on the raw data. Some of the Wireless technologybased sensor data are lost during the propagation. To resolve this problem, linear interpolation algorithms are implemented. After this, scaling and normalization are done on the data to avoid training bias [3]. Execute a sliding window for sensor data segmentation. And this segmented data is given to LSTMCNN architecture. Each LSTM layer has 32 neurons and each one is used to extract the time-dependent features of the data sequence. Different gates of the LSTM [3] control the 32 memory cells of the LSTM. To achieve this control mechanism, input data sequences are fed into the different gates of the LSTM. LSTM output dimensions are different from the input dimension of CNN. To equate this dimension, the dimension of the output of the LSTM is modified.

Rectified Linear Unit (ReLU) is used in CNN to obtain the distinct feature map.

$$b_{jk} = f\left(\sum_{l=1}^{L}\sum_{m=1}^{M} V_{l,m}x_{j+lk+m} + c\right) \qquad (3)$$

where $b_{jk}$is the corresponding activation, $V_{l,m}$denotes the $l\times m$weight matrix of convolution kernel, x_{j+lk+m} indicates the activation of the upper neurons connected to the neuron (j,k), c is the bias value, and f is a non-linear function.

The size of each convolution layer is provided in Fig 3. Not at all like the fully connected layer, the global average pooling layer plays out a worldwide averaging pooling activity on each element map. During the backpropagation in the

training period of the neural network, each weight is refreshed. This makes some training issues like a decrease in the training speed, convergence, etc. The batch normalization layer is placed after the GAP layer. Batch normalization could increase the training rate because of the normalization of the inputs.In the output layer, the fully connected layer output vector is given to the softmax classifier. And the softmax classifier classifies the particular human actions by the probability distribution. This method is efficient to simplify the feature extraction.

D. Temporal Segment Connection Network for Action Recognition

The High presentation of two-stream Convolutional Neural Networks in video human activity recognition, it has been utilized more in human activity recognition. Most existing works train each testing group independently, or just epitomized at the last stage, which disregards the progression of activity in temporal and the integral data between activity pieces. In this strategy for activity recognition, a transient portion association network is proposed to conquer these impediments see Fig 4[4]. In this organization, the failure to remember the forget gate module of the long short-term memory (LSTM) network is utilized to build up component level associations between each testing gathering. Then again, a bi-directional long short-term memory (Bi-LSTM) network is utilized to naturally assess the significance loads of each inspecting bunch dependent on the profound element grouping. The trial results on UCF101[11] and HMDB51[10] datasets show that the proposed model can successfully improve the use pace of temporal information and the capacity of general activity portrayal, consequently essentially improves the exactness of human activity identification.

Temporal Segment Connection Network has made the following contributions:

- *The feature-level forget-gate connections are set up between contiguous testing groups, which can improve the temporal connection, yet additionally, extricate the integral data between the inspecting input group samplings.*
- *The system of enriching weights dependent on the setting data is presented, which can consequently assess the significance weights of each sampling group.*
- *The model achieves a promising performance in recognizing actual human activity on two reference data sets, including UCF101 and HMDB51 individually.*

E. An Accurate Device-Free Action Recognition System Using Two-stream Network

With the promotion of Wi-Fi signals and pressing requests for latent human activity acknowledgment, remote detecting based movement acknowledgment has been an intriguing issue lately. Most existing investigations depend on customary handmade highlights and restricted work centers around how to viably separate deep features with spatial-temporal information. In this method, build up an exact gadget-free activity acknowledgment framework using a Commodity Off-The-Shelf (COTS) switch and propose a novel profound learning system (named two-stream network) mining spatial-temporal prompts in channel state information (CSI). In particular, a whole activity test is sectioned into an arrangement of intelligible sub-action cuts. At that point the method attempt to catch the integral highlights on appearance from the first CSI clasps and movement between CSI outlines. The spatial and transient data are prepared with independent organizations which are at that point coordinated for the last acknowledgment task. The broad tests are actualized on the information gathered from two indoor conditions, separately arriving at 97.6% and 96.9% accuracy.

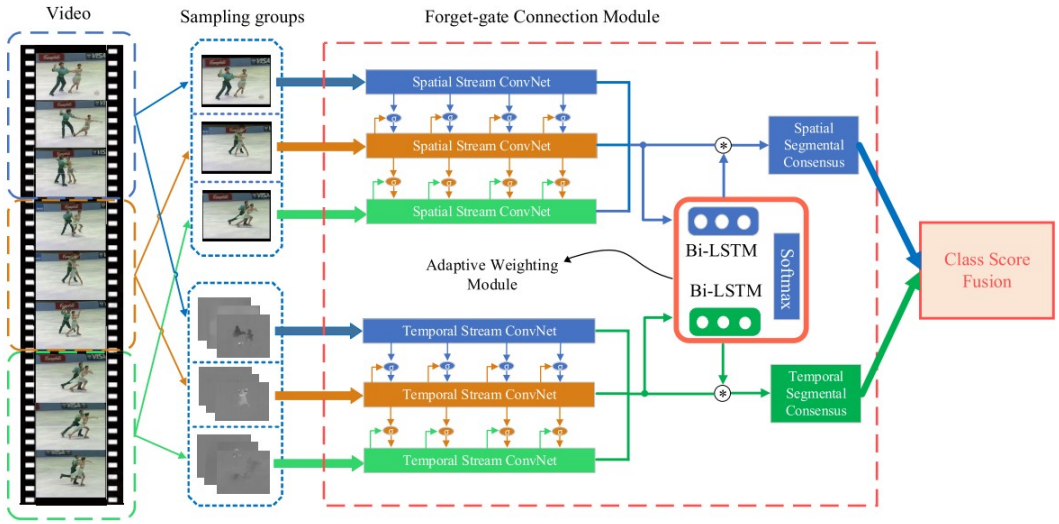

Fig. 3. Temporal Segment Connection Network (TSCN)[4]

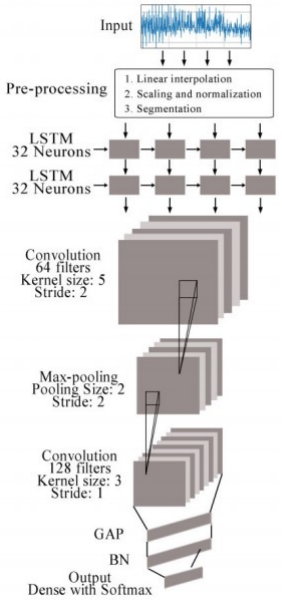

Fig. 4. Framework scheme of the LSTM-CNN model.[3]

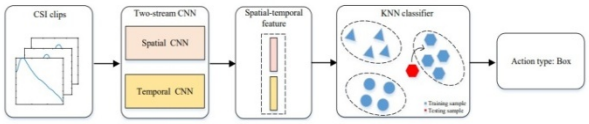

Fig. 5. The process of feature extraction and classification for an action sample.[5]

The fundamental test behind this technique is that frames are equivalent size input test for CNNs, since the time length of channel state data isn't the equivalent consistently. This technique is an activity recognition strategy without utilizing LSTM, so the Spatio-temporal extraction is additionally a test in this technique. To tackle this difficult situation two-stream convolutional neural organization is utilized. Both the temporal and spatial streams are remembered for one CNN architecture. The contributions of the spatial stream are the sufficiency appearance of the CSI and for the spatial stream, it is the distinction of the adequacy of the CSI. The two streams are actualized on planned multi-layer CNN structure and the last completely associated layer that is the fully connected layer of each stream is joined as the spatio-temporal feature includes with classifier, for example, K-Nearest Neighbors(KNN) followed, see Fig 5[5].

The CSI contains high-frequency parts from the noise during the signal propagation. So, these high-recurrence segments are filtered by utilizing a Butterworth filter during information processing. More definitely, here the amplitudes of the CSI are communicated as follows with received package number N:

$$H_A = (h_1, h_2, \ldots, h_N) \subseteq R^{90 \times N} \qquad (4)$$

As a result of the time span between two nonstop activities, one of the two beginning stages (or endpoints) whose separation is excessively close will be considered as a false positive.

IV. APPLICATIONS

A.Automation Application

A Human Activity Recognition framework which shapes a section of the video reconnaissance framework was planned and executed. The acknowledgment framework was planned based on the Human skeletal highlights which were collected utilizing a Kinect sensor. The acknowledgment framework was created as a product framework that takes contributions from the Kinect sensor and conveys the location result in an Automation framework. The Automation framework was planned as a mechanization framework with four gadgets controlled for four various activities from the client. The framework was effectively planned and actualized on Matlab along with the

fundamental equipment and programming assets.

B. Multimodal systems in the medical domain

A multi-sensor stage that screens and perceives exercises of an individual utilizing tangible gadgets conveyed at various purposes of the body[14][15]. It distinguishes the client exercises progressively and records these characterization results during the day [8]. At that point, the different time-spaces include sets and inspecting rates are contrasted with break down the compromise between acknowledgment precision and computational multifaceted nature. Order precision is determined for each movement gathered from six distinctive body focuses (wrist, pocket, pack, neckband, shirt, and belt). The information from all subjects is at last consolidated to prepare the overall classifier. Because of the performed assessment, the gathered information showed that each of the six focuses was useful for distinguishing strolling, standing, sitting, and running exercises.

C. Visual systems for security and surveillance systems in public areas

Security and observation frameworks have utilized visual handling approaches broadly to follow human practices in open conditions. Visual sensor-based strategies[16] are the most appropriate methodologies for actualizing such frameworks because of the substantial evidential verifications which can be given by recordings and pictures because of their inclination.

V. CONCLUSION

While considering large variations and complexity of body postures, occlusion, background noise, and camera movements, recognition of human activities remains a standard problem in machine vision or in the understanding of video data. To find actions in a particular input like images and videos, the first step is to select appropriate data for capturing the action. Furthermore, a significant algorithm must be used for recognizing human action. For human activity feature understanding and extraction, deep neural networksbasedmethods have good performance. Apart from the classification of primary and individual actions, the recognition of interactions of humans with objects and the detection of actions has become the new leading research topics.

REFERENCES

[1] Heng Wang, Alexander Klaser, Cordelia Schmid, Liu Cheng-Lin. Action¨ Recognition by Dense Trajectories. CVPR 2011 - IEEE Conference on Computer Vision Pattern Recognition, Jun 2011,Colorado Springs, United States. pp.3169-3176, ff10.1109/CVPR.2011.5995407ff.

[2] Lei Wang, Yangyyang, Jun Cheng, Haiying,Jianqin, and Jiaji, (Member, IEEE), ”Human Action Recognition by Learning Spatio-Temporal Features With Deep Neural Networks”, IEEE Transactions on computer vision 2018, Volume: 6, pp.17913 - 17922, doi: 10.1109/ACCESS.2018.2817253

[3] Qian Li , Wenzhu Yang, Xiangyang Chen,Tongtong Yuan,and Yuxia Wang, ”LSTM-CNN Architecture for Human Activity Recognition”, IEEE Access Volume: 8, pp.56855 - 56866, 2020, doi:10.1109/ACCESS.2020.2982225

[4] Qian Li , Wenzhu Yang, XiangyangChen,Tongtong Yuan, AND Yuxia Wang, ”Temporal Segment Connection Network for Action Recognition”, IEEE Access Volume: 8, pp.179118 - 1791272020, 2020, doi:10.1109/ACCESS.2020.3027386

[5] Biyun Sheng, Yuanrun Fang, Fu Xiao, Lijuan Sun, ” Accurate Device-Free Action Recognition System Using Two-stream Network”, IEEE Transactions on Vehicular Technology,pp. 7930 - 7939, 2020, doi.10.1109/TVT.2020.2993901

[6] N. Ikizler-Cinbis and S. Sclaroff, ”Object, scene and actions: Combining multiple features for human action recognition”, In ECCV, 2010.

[7] J. Y. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, “Beyond short snippets: Deep networks for video classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR),Jun. 2015, pp. 4694–4702.

[8] S. Sharma, R. Kiros, and R. Salakhutdinov. (2015). “Action recognition using visual attention.” [Online]. Available: https://arxiv.org/abs/1511.04119

[9] 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.

[10] H. Kuehne, R. Stiefelhagen, T. Serre, and H. Jhuang, “HMDB51: A large video database for human motion recognition,” in High Performance Computing in Science and

Engineering. Berlin, Germany: Springer,2013, pp. 571–582.

[11] 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

[12] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In ECCV, 2006

[13] Maurer, U, Smailagic, A, Siewiorek, DP. Activity recognition and monitoring using multiple sensors on different body positions. In: BSN 2006, international workshop wearable and implantable body sensor networks, Cambridge, MA, 3–5 April 2006, p.4. New York: IEEE.

[14] Yongbin Gao, Xuehao Xiang, NaixueXiong, Bo Huang, Hyo Jong Lee, Rad Alrifai, Xiaoyan Jiang, Zhijun Fang, "Human Action Monitoring for Healthcare Based on Deep Learning" IEEE Access, Volume. 6,pp. 52277 - 52285,2018, doi: 10.1109/ACCESS.2018.2869790

[15] Kiran Talele, KushalTuckley, "Human Action Unit detection of patient using geometric feature analysis" in Proc.2016 IEEE Region 10 Conference (TENCON) Nov. 2016, doi: 10.1109/TENCON.2016.7848408

[16] G.L. Foresti, C. Micheloni, L. Snidaro, P. Remagnino, T. Ellis, "Active video-based surveillance system: the low-level image and video processing techniques needed for implementation" , IEEE Signal Processing Magazine, Volume: 22, pp.25 - 37, 2005, doi:10.1109/MSP.2005.1406473