ISSN: 2366-1313

HUMAN ACTION RECOGNITION USING DEEP LEARNING BASED CNN MODEL

[1]V. NAVEEN KUMAR, [2]DONDETI RAM MOHAN REDDY

[1]MTech Student, Dept. of CSE, Newton's Institute of Engineering, Guntur, (A.P)

[2]Associate professor, Dept. of CSE, Newton's Institute of Engineering, Guntur, (A.P)

Abstract*: The goal of human action recognition is to identify and understand the actions of people in videos and export corresponding tags. In addition to spatial correlation existing in 2D images, actions in a video also own the attributes in temporal domain. Due to the complexity of human actions, e.g., the changes of perspectives, background noises, and others will affect the recognition. In order to solve these thorny problems, three algorithms are designed and implemented in this paper. Based on convolutional neural networks (CNN), Two-Stream CNN, CNN+LSTM, and 3D CNN are harnessed to identify human actions in videos. Each algorithm is explicated and analysed on details. HMDB-51 dataset is applied to test these algorithms and gain the best results. Experimental results showcase that the three methods have effectively identified human actions given a video, the best algorithm thus is selected.*

Keywords*: Deep learning, convolutional neural network, LSTM, human action recognition*.

I. INTRODUCTION

Human activity recognition (HAR) is a well-known research topic, that involves the correct identification of different activities, sampled in a number of ways. In particular, sensor-based HAR makes use of inertial sensors, such as accelerometers and gyroscopes, to sample acceleration and angular velocity of a body. Sensor-based techniques are generally considered superior when compared with other methods, such as vision-based, which use cameras and microphones to record the movements of a body: they are not intrusive for the users, as they do not involve video recording in private and domestic context, less sensitive to environmental noise, cheap and efficient in terms of power consumption [8, 13]. Moreover, the wide diffusion of embedded sensors in smartphones makes these devices ubiquitous. One of the main challenges in sensor-based HAR is the information representation. Traditional classification methods are based on features that are engineered and extracted from the kinetic signals. However, these features are mainly picked on a heuristic base, in accordance with the task at hand. Often, the feature extraction process requires a deep knowledge of the application domain, or human experience, and still results in shallow features only [5]. Moreover, typical HAR methods do not scale for complex motion patterns, and in most cases do not perform well on dynamic data, that is, data picked from continuous streams. On this regard, automatic and deep methods are gaining momentum in the field of HAR. With the adoption of data-driven approaches for signal classification, the process of selecting meaningful features from the data is deferred to the learning model. In particular, CNNs have the ability to detect both spatial and temporal dependencies among signals, and can effectively model scale invariant features. In this paper, we apply convolutional neural networks for the HAR problem. The dataset we collected is composed of 16 activities from the Otago exercise program. We train several CNNs with signals coming from different sensors, and we compare the results in order to detect the most informative sensor placement for lower-limb activities. Our findings show that, in most scenarios, the performance of a single sensor is comparable to the performance of multiple sensors, but the usage of multiple sensor configurations yields slightly better results. This suggests that collinearities exist among the signals sampled with sensors on different placements[1].

Human action recognition from videos is based on the analysis of a sequence of video frames by using computers, so as to automatically find human actions without manual operations. In the era of the Internet with mobile phones, people’s daily lives have been surrounded by access control such as building gates, traffic sensors, security cameras, and many others. The ubiquitous cameras enable everyone's actions in public to be monitored, identification of human actions in surveillance videos has tremendous significance in the field of cybersecurity. In addition, analysis and understanding of human actions in digital videos encapsulate multiple interesting research topics such as object detection, semantic segmentation, motion analysis, etc. Hence, human action analysis has a broad spectrum of applications including intelligent surveillance, intelligent care, etc. Traditional methods in machine learning for human action recognition extracted visual features primarily based on human observations. It is subject to a vast amount of human experience and background knowledge. Most of these algorithms only performed well on the exact dataset for a specific experiment. At present, there are a huge number of digital video footages available on the Internet, like YouTube, it is impossible to satisfy the demand to annotate all videos with tags and extract the features only based on our human labour. Fortunately, the surge of deep learning methods in recent years has provided a solution. Deep learning algorithms generate feature maps based on artificial neural networks [2]. Deep neural networks have remarkable achievements in the field of computer vision, natural language processing, robotics. However, as deep learning is still at its early stage. Meanwhile, human motion is relatively complicated, the relevant motion analysis is affected by various determinants such as chaotic background, various lighting conditions, unstable image acquisition, and insufficient pattern classes.

II. LITERATURE SURVEY

In recent years, a great deal of breakthroughs has been attained in the field of machine vision and deep learning. A plenty of methods for human action recognition based on deep learning have been explored and exploited [3]. Compared with machine learning methods for human behaviour recognition, deep learn approaches do not require a specific type of human experience and knowledge. Instead, human actions in a video are identified directly in the end-to-end way [4]. According to feature extraction methods, the approaches are grouped into two categories, i.e., human action recognition based on skeletons, human action recognition based on feature maps. Amongst the deep learning methods, spatiotemporal networks and Two-Stream networks are the salient ones [5]. In these methods, CNN and RNN are most popular [6]. A multimodal learning approach was proposed for the recognition and classification of human actions [6,29]. In 2017, 3D convolutional neural network (3D CNN) and twoway long short-term memory network (ConvLSTM) were trained based on multimodal and spatiotemporal data to fulfil human action recognition by using support vector machine (SVM). A deep dynamic neural network (DDNN) was designed to implement action recognition from input data under multimodal framework, which extracts spatiotemporal features from RGB and RGB-D images. A scene-flow dynamic model was deployed to generate visual features from RGB and depth images, which were imported for training by using CNN networks. A 3D deep convolutional neural network was taken into account to learn high-level features from the original images, fix the position and angle of bone joint information. The two features were fused by using SVM for human action classification. In 2018, CNN and RNN were integrated together to cope with the spatiotemporal information of human actions and achieved promising result.

A similar technique is suggested by Yang et al. [7]. In their work, they use the same public datasets, however, they apply 2-D convolution over a single-channel representation of the kinetic signals. This particular application of CNNs for the activity recognition problem is further elaborated by Ha et al. [8], with a multi-channel convolutional network that leverages both acceleration and angular velocity signals to classify daily activities from a public dataset of upper-limb movements. The classification task they perform is personalized, so the signals gathered from each participant are used to train individual learning models.

III. PROPOSED SYSTEM

The ultimate goal of this paper is to implement human action recognition from the given videos. We segment the video footages and imported the video frames as the input data. Three deep learning methods are applied to generate feature maps for human action recognition. Throughout network training, we recognize human actions and finally export the class tags.

CNN+LSTM Model

CNNs are a class of feedforward neural networks, which are principally comprised of input layer, convolutional layer, pooling layer, full connection layer, and output layer. The convolutional layer of a CNN encompasses one or more feature planes. Each feature plane is related to numerous neurons in a region, the neurons in the same plane share the same weights. The shared weights consist of network parametric set, the better weights are gained in the process of model training. By extracting local features and synthesizing them at a higher level, CNNs not only yield global features but also lessen a number of neuron nodes. At this point, the number of neurons is still very large, by setting the weight for each neuron equally, the number of network parameters will be greatly diminished. On the first convolution layer, the output is �� then the output after � times of convolution operations is

$$y_k^m = \delta(\Sigma_{y_i^{n-1} \in M_k} y_i^{m-1} * W_{ik}^m + b_k^m)$$

where �(.) is an activation function, �� is based on a layer of feature collection, ��� refers to convolution kernels, ∗ means a convolution, �� stands for offset. In CNNs, pooling layer follows the convolutional layer to reduce dimensionality and accelerate the convergence of network training. The other is to remove redundant features so as to prevent overfitting. Each neuron in the full connection layer is linked with all neurons in the proceed layer. Throughout the full connections, all local features are integrated together to form the overall features. Each neuron in the full connection layer operates with an activation function, which is transferred to the output layer. In RNNs, the memory units have not ability to measure the value of information. It is impossible to distinguish the importance of state information, which results in the useless information being stored in memory. However, the truly valuable information is squeezed. Each unit of LSTM network contains memory unit, input gate, forget gate, and output gate.

Behavioral videos contain not only spatial data but also temporal information. Using CNN, the temporal information of a given video cannot be fully used. The output of LSTM is determined by using combined actions of current input and historical output. Temporal information is applied to represent a sequence of video frames. The structure for CNN+LSTM model is shown in Fig. 1.

n CNN, a video is decomposed into single frames so as to form a large image dataset. This set is imported as the input of single-channel CNN+LSTM for pretraining. The training results are stored and the sequence of features are generated. The dataset is then imported into LSTM network as input data. The sequence of video frames is used to train LSTM network After the training, the parameters of CNN are exported as spatial features for human action recognition

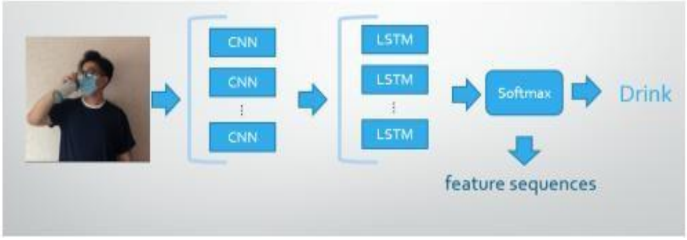

Fig.1 The structure of CNN+LSTM algorithm

In each video sequence, every 8 video frames are treated as a group of input. The spatial feature is imported into the LSTM to learn the temporal relationship of the frame sequence so as to fix the network parameters. In the test, every 8 video frames extracted from a video equally are taken as the input data of CNN+LSTM model. After spatial feature extraction and temporal feature selection, the output tags

ISSN: 2366-1313

of LSTM are thought as the final classification result

IV. ANALYSIS AND DISCUSSIONS

The advantage of the Two-Stream CNN model is that two convolutional neural networks are able to acquire pretty rich features of human actions. In the model, the diversity of behavioural description is increased and more determinants are obtained. But CNN is limited by its own attributes, the Two-Stream CNN is apt. This structure ignores the temporal relationship of the video frames. It is difficult to cope with video samples with complicated spatiotemporal relationships and abrupt changes. In addition to optimize the performance of the CNN, we should consider modelling the temporal information of video frames such as adding more temporal convolution operations. 3D CNN increases the dimensionality of data input even if the training samples are not raised. Therefore, 3D CNN needs more samples of human actions to train the network well. Under the same training condition, the recognition accuracy is not excellent. In addition, 3D CNN only copes with adjacent frames if short-term motion information of the actions is available. This method is still unable to model the full video sequence. In time series analysis, 3D convolutions are offered to enrich 2D convolutions, a basic network is employed to extract spatial features and short-time motion features for complicated action recognition. CNN+LSTM model makes up the CNN for recognizing human actions. Human action recognition benefitsfrom visual information of digital videos and the temporal relationship between video adjacent frames. Although the architecture of CNN+LSTM is clear, what information the structure needs is still ambiguous. By effectively integrating multimodal actions, we are able to learn from each other and gain much discriminative action descriptors. Overall, all three methods accurately identified human actions. More accurate identification is attained by adjusting the parameters of these nets. In our experiments, though CNN+LSTM has the best performance, the Two-Stream CNN and 3D CNN have shown better outcomes and will be continuously to be improved.

V. CONCLUSION

Feature extraction is the most critical step in human action recognition, which relies on domain knowledge and human experience that cannot meet the demands of data growth. Therefore, we take deep learning methods as our start point in this paper. The mainstream method in deep learning is based on CNNs. Therefore, three recognition algorithms, i.e., the Two-Streams CNN, CNN+LSTM, and 3D CNN are chiefly taken into account in this paper. Throughout feature selection, the algorithms successfully recognized human actions from a given video, they are distinct in dealing with time series problems. Our experiments show that LSTM better deals with this temporal coherence. Therefore, LSTM+CNN recognizes human actions from the videos more effectively.

REFERENCES

1. Alsheikh, M.A., Selim, A., Niyato, D., Doyle, L., Lin, S., Tan, H.P.: Deep activity recognition models with triaxial accelerometers. CoRR abs/1511.04664 (2015), http://arxiv.org/abs/1511.04664

2. Banos, O., Galvez, J.M., Damas, M., Pomares, H., Rojas, I.: Evaluating the effects of signal segmentation on activity recognition. In: International Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO 2014. pp. 759–765 (2014) 4. Bengio, Y.: Practical recommendations for gradient-based training of deep architectures. CoRR abs/1206.5533 (2012), http://arxiv.org/abs/1206.5533

3. Bengio, Y.: Deep learning of representations: Looking forward. CoRR abs/1305.0445 (2013), http://arxiv.org/abs/1305.0445 6. Bulling, A., Blanke, U., Schiele, B.: A tutorial on human activity recognition using body-worn inertial sensors. ACM Comput. Surv. 46(3), 33:1–33:33 (Jan 2014). https://doi.org/10.1145/2499621, http://doi.acm.org/10.1145/2499621 4. Burns, A., Greene, B.R., McGrath, M.J., O'Shea, T.J., Kuris, B., Ayer, S.M., Stroiescu, F., Cionca, V.: ShimmerTM a wireless sensor platform for noninvasive biomedical research. IEEE Sensors Journal 10(9), 1527 – 1534 (2010). https://doi.org/10.1109/JSEN.2010.2045498

5. Cook, D., Feuz, K.D., Krishnan, N.C.: Transfer learning for activity recognition: a survey. Knowledge and Information Systems 36(3), 537–556 (Sep 2013). https://doi.org/10.1007/s10115-013-0665-3, https://doi.org/10.1007/s10115-013- 0665-3

6. S. Khan, H. Rahmani, S. Shah, M. Bennamoun, G. Medioni, S. Dickinson, A Guide to Convolutional Neural Networks for Computer Vision, Morgan & Claypool, 2018.

7. L. Jing, Y. Ye, X. Yang and Y. Tian, 3D convolutional neural network with multi-model framework for action recognition, IEEE International Conference on Image

Processing (ICIP), Beijing, 2017, pp. 1837-1841, doi: 10.1109/ICIP.2017.8296599.

8. C. Li, S. Sun, X. Min, W. Lin, B. Nie and X. Zhang, End-to-end learning of deep convolutional neural network for 3D human action recognition, IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Hong Kong, 2017, pp. 609-612.

9. D. Wu et al., “Deep dynamic neural networks for multimodal gesture segmentation and recognition, ” IEEE Transactions on Pattern Analysis and Machine Intelligence, 38(8) 1583-1597, 2016.