AMERICAN Journal of Engineering, Mechanics and Architecture
Volume 2, Issue 4, 2024 ISSN (E): 2993-2637

Identifying Human Actions using Convolutional Neural Networks

Abdurashidova Kamola Turgunbaevna
Associate professor at Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Department of Computer Systems

Abdukhakimov Fayzulla Kudratulla ugli
Graduate student at Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Department of Computer Systems

Chorshanbiyeva Sevinch Akramovna
Undergraduate student at Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Department of Computer Systems

Abstract: Video indexing, intelligent surveillance, multimedia understanding, and other domains all make extensive use of video action recognition. Lately, it was significantly enhanced by adding deep learning through Convolutional Neural Network (CNN) learning. This inspired us to examine the noteworthy efforts on action recognition using CNN. This paper presents a clear and objective overview of CNN-based action recognition and offers recommendations for further research.

Keywords: CNN, feature engineering, spatial-temporal activity, poses, pooling.

INTRODUCTION

Since it is crucial to robotics, human-computer interaction, intelligent surveillance, and other fields, the recognition and comprehension of human behaviors and intentions has been a significant and well-liked research issue. Many approaches for video action recognition have been proposed in the last few decades. Furthermore, a number of action recognition datasets and benchmarks have been made public. Since their widespread use in image analysis in 2012, convolutional neural networks (CNNs) [1] have significantly improved tasks including object detection [2], scene categorization [3], and image classification [4]. The use of CNNs to identify actions in videos has gained traction as a result of this success. A thriving research topic centered around CNN-based action recognition has resulted from substantial breakthroughs in techniques and CNN structures optimized for video. In this area, a plethora of CNN-based methods have surfaced, showing impressive performance recently.

Video action recognition involves two primary objectives, as illustrated in Figure 1 [5]: classifying a video into established action classes and determining the temporal occurrence of predefined actions within a video. Classification and detection are the popular terms for these activities, respectively.

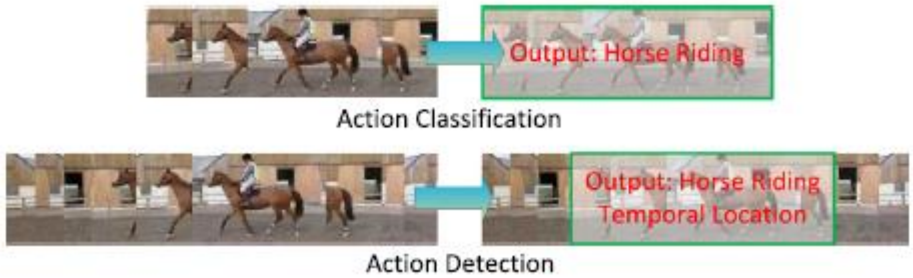

Figure 1. Two categories of action recognition task

Depth video can capture the geometric intricacies of objects and is not impacted by changes in lighting, unlike visible (gray or RGB) video. In a similar vein, infrared thermal video is resilient to changes in illumination and difficult lighting situations. Furthermore, the use of multi-view cameras makes it possible to obtain more thorough data from various angles, which improves action detection accuracy. Consequently, action detection using depth [6], infrared [7], and multi-view video [8] has attracted a lot of attention. Moreover, activities in multi-view, infrared, and depth video have recently been recognized using Convolutional Neural Networks (CNNs).

It is clear that CNN-based action recognition can analyze not only single-view but also multi-view movies, as well as depth and infrared videos in addition to visible videos. The detection job is intrinsically interwoven with classification, an important and well-studied task. Therefore, the task of action classification in single-view viewable videos is the focus of this paper.

METHODOLOGY

The Convolutional Neural Network (CNN) is a type of feedforward artificial neural network, inspired by biological processes. It comprises an input layer, an output layer, and multiple hidden layers, which can be convolutional, pooling, or fully connected. In a convolutional layer, the network applies a convolution operation and adds a bias to the input data. The result is passed through an activation function before being forwarded to the next layer. The convolution operation at a specific position (x, y) in the j-th feature map in the i-th layer is mathematically described by Equation (1).

$$C_{il}^{xy} = \varphi(b_{i,j} + \sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} w_{i,j,m}^{p,q} v_{(i-1),m}^{(x+p),(y+p)}) \qquad (1)$$

A weight matrix (w) and a kernel with dimensions P (height) and Q (width) make up the output of the convolutional layer, to which the activation function—such as Tanh, Sigmoid, or ReLU—is applied. Downsampling is done non-linearly by the pooling layer. After these layers, the CNN uses fully linked layers—where every neuron is coupled to every activation from the layer before it—to do high-level reasoning.

The original goal of the CNN model LeNet-5 [1], which was developed in 1998, was to identify numbers in documents. But until it achieved a breakthrough in image classification [4] in 2012, its growth was slow. Significant progress has been made in object detection and picture categorization using CNN technology throughout the years. Several CNN architectures have been developed, including ZFNet [9], VGG [10], GoogLeNet [11], BN-Inception [12], and ResNets [13]. These designs rely on large-scale datasets for training to derive their pre-trained models, or weights. To improve the previously learned network models, extra training, or transfer learning, is frequently carried out when working with novel small-scale datasets or diverse kinds of data. Motivated by CNN's achievements in image processing, scientists have also utilized CNN methods for video action identification. An increasing variety of CNN-based techniques for action recognition have demonstrated strong results. CNNs are typically used in 2D space and are particularly good at extracting spatial features from static images, as shown by Equation (1). CNNs have been used in some action recognition research to extract spatial information, which is used in conjunction with handmade characteristics like iDT to achieve final action recognition. CNNs were initially intended for the extraction of spatial features, not temporal awareness, which makes them less appropriate for movies, which are intrinsically 3D spatiotemporal signals. Therefore, using temporal information is the key to expanding CNNs from images to movies. We group temporal information-exploitation options into three categories: 1) 3D convolution; 2) motion-related data incorporation as CNN input; and 3) fusion. These tactics frequently intersect, like in the case of using motion-related data as input for a 3D CNN architecture. Based on these approaches, this section will look at CNN-based action recognition techniques.

Using 3D convolution on movies is a straightforward way to leverage spatiotemporal information; this strategy has been verified in early CNN-based action recognition research conducted before 2012. Convolution in 3D is the process of convolving a video clip using a 3D kernel. Equation (2) provides a formal definition of the mathematical operation at point (x, y, z) in the j-th feature map in the i-th layer.

$$C_{il}^{xy} = \varphi(b_{i,j} + \sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} \sum_{r=0}^{R_i-1} w_{i,j,m}^{p,q,r} v_{(i-1),m}^{,(y+p)(z+r)}) \qquad (2)$$

The non-linear activation function, ϕ, in this equation can be Tanh, Sigmoid, or ReLU. The 3D weight matrix is represented by w, while the kernel's height, width, and temporal length are indicated by P, Q, and R, respectively. Figure 2 shows the steps involved in 2D and 3D convolution.

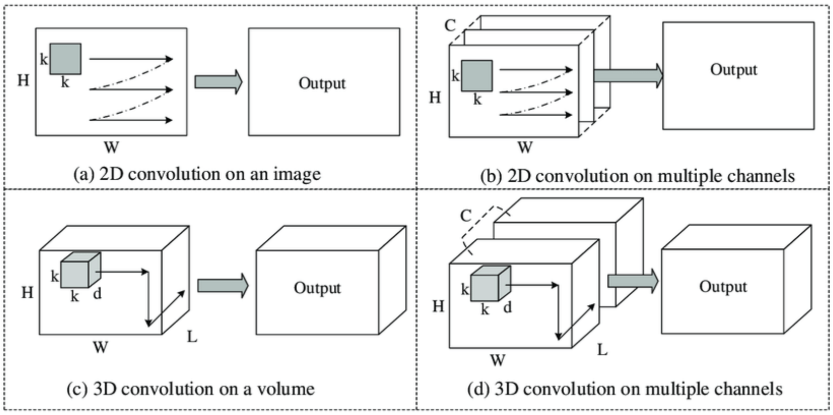
k
k
H
W
Output
(a) 2D convolution on an image
C
k
k
H
W
Output
(b) 2D convolution on multiple channels
k
k
d
H
W
L
Output
(c) 3D convolution on a volume
C
k
k
d
H
W
L
Output
(d) 3D convolution on multiple channels

Figure 2. Comparison of 2D convolution and 3D convolution

An important early work on action recognition was presented prior to 2012 by Ji et al. [14], who created a 3D CNN architecture with one fixed layer, three convolutional layers, two subsampling layers, and one fully connected layer. Gray, gradient, and optical flow channels are created by the fixed layer, and each channel is then subjected to subsampling and convolution processes. Information from all channels is integrated to generate the final action representation. By incorporating predictions from multiple 3D CNN architectures and regularizing outputs with high-level features, Ji et al. [15] improved their original 3D CNN model.

DISCUSSION AND RESULTS

An overview of CNN-based action recognition techniques based on temporal information leveraging strategies was given in this section. It is imperative to employ two-stream techniques and 3D CNN in order to efficiently capture spatiotemporal data. Fusion, as used in CNN-based action recognition, is a more general term that refers to the process of merging, combining, or aggregating various extracted information types in order to take use of spatiotemporal cues. Additionally, there are four temporal scales into which the spatiotemporal information that CNN retrieved can be divided: apparent (spatial) information, motion information, short-term temporal information, and long-term temporal information. CNNs handle short-term temporal information on brief video clips, optical flow information on motion, and apparent information on individual frames. LSTM can be used to process extracted deep features, long-term temporal 3D convolutions can be used for longer video clips, and identity mapping kernels can be integrated as temporal filters to handle long-term deep temporal information. As action recognition advances, various benchmarks and datasets have been introduced. Hassner et al. [16] conducted a review of these datasets, categorizing them into early datasets collected in laboratory settings, such as Weizmann [17]; interim datasets sourced from television footage, like UCF Sports [18]; and recent datasets captured in real-world environments, such as UCF101 [19]. In this section, we present several recent large-scale action recognition datasets collected from real-world scenarios.

HMDB51 [85]: The videos within HMDB51 were sourced from diverse internet platforms and digitized films, showcasing human actions typical of daily life. Challenges in this dataset include significant variations in camera angle and movement, cluttered backgrounds, and fluctuations in actor positions, scale, and appearances. HMDB51 encompasses 51 distinct action categories, each comprising a minimum of 101 clips, totaling 6766 video clips.

UCF101 [84]: Comprising 13,320 films from 101 action categories on YouTube, UCF101 is an expansion of the UCF50 [86] dataset. With notable variances in camera movement, item appearance and position, object scale, viewpoint, crowded backdrops, lighting conditions, and other aspects, it offers the widest variety of actions. Each action category has twenty-five groups of videos, each of which has four to seven videos that each show a different activity.

Relatively few action recognition studies have been assessed on the Sports-1M dataset due of its wide breadth. We provide the results of CNN-based techniques on Sports-1M in Table 1. Nonetheless, the most popular benchmarks for assessing modern action detection methods are HMDB51 and UCF101. We provide a thorough overview of CNN-based strategies that have produced noteworthy effects or outcomes on these two datasets in Table 2. We outline the CNN architecture, the recognition accuracy, and the input that each strategy uses. We incorporate the findings of iDT-related action recognition [24,25] to compare with the state-of-the-art manual approaches. The original studies are the source of the reported accuracy. Furthermore, we have highlighted in red the state-of-the-art CNN-based methods.

Table 1. Recognition results on sports-1M.

Method	Hit@5 (%)	CNN architecture	Input of the CNN
SlowFusion	*80.1*	*AlexNet*	*RGB*
LSTM	*90.2*	*GoogleNet*	*RGB*
HumanSkeleton	*84.6*	*GoogleNet*	*RGB*
P3D-ResNet	*88.4*	*P3D-ResNet*	*RGB*
C3D	*84.9*	*C3D*	*RGB*

A quick glance at Table 2's results shows that, whereas action recognition accuracy on UCF101 surpasses 94%, it remains approximately 70% on HMDB51. Issues including substantial viewpoint variance, crowded backgrounds, and shifts in actor placements, scale, and looks are to blame for HMDB51's inferior accuracy. This implies that these challenges are too great for the action recognition techniques used today. Furthermore, iDT-related techniques outperformed earlier CNN-based algorithms. However, recently developed CNN-based approaches have demonstrated tremendous improvement and have outperformed iDT-related methods with the use of deeper CNN architectures and new technology.

Method	UCF101 mAP(%)	HMDB51 mAP(%)	CNN architecture	Input of the CNN
TDD	89.7	63.2	ZFNet	RGB+OF
Two-stream SVM	88.2	59.1	CNN-M	RGB+OF(RGB)
HRP	91.4	66.9	VGG-16	RGB
ActionVLAD	92.6	66.8	VGG-16	RGB+OF(RGB)
AdaScan	89.2	54.9	VGG-16	RGB+OF(RGB)

GRP	91.1	65.4	VGG-16	RGB+OF
Spatio-temporal LSTM	83.0	55.2	C3D	RGB+OF
LTC	91.6	64.8	LTC	RGB+OF(RGB)
Res3D	85.8	54.9	Res3D	RGB
ST-VLMPF(DF)	93.6	69.5	VGG-16, VGG-19, C3D	RGB+OF
SSN	94.8	73.8	BN-Inception	RGB+OF

CONCLUSION

The computer science community has recently paid close attention to CNN-based action recognition because of its outstanding performance, which outperforms handcrafted representation techniques on difficult datasets. As demonstrated in the image domain, the capacity to efficiently extract spatial information from 2D space fits in nicely with the intuitive understanding of video as a 3D spatiotemporal signal. The main difficulty in using CNNs on video is utilizing temporal information. This study offers a thorough overview of methods for CNN-based action recognition that take advantage of temporal information. Furthermore, we showcase and contrast the outcomes of CNN-based action detection techniques on two difficult datasets: HMDB51 and UCF101. In order to further enhance action recognition, we anticipate multiple avenues of future study for CNN-based techniques. First off, when it comes to learning spatiotemporal characteristics, 3D CNNs outperform 2D CNNs. Consequently, it makes sense to convert the potent architectures created for 2D CNNs to 3D CNNs. Although employing optical flow as input has improved CNN-based action recognition, its computing needs remain high. Therefore, it is crucial to investigate fresh, effective motion-related data for CNN input. Furthermore, fusion will remain essential to CNN-based action recognition since it allows spatiotemporal information to be used by combining, pooling, or aggregating different kinds of extracted data.

REFERENCES

1. Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE (1998) 2278–2324.
2. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for ac- curate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 580–587.
3. C. Farabet, C. Couprie, Y. LeCun, Learning hierarchical features for scene labelling, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1915–1929.
4. A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS), 2012, pp. 1097–1105.
5. H. Idrees, A. Zamir, Y. Jiang, A. Gorban, I. Laptev, R. Sukthankar, M. Shah, The THUMOS challenge on action recognition for videos “in the wild”, Computer Vision Image Understanding 155 (2017) 1–23.
6. R. Yang, R. Yang, DMM-pyramid based deep architectures for action recognition with depth cameras, in: Proceedings of the Asian Conference on Computer Vision (ACCV), 2014, pp. 37–49.
7. C. Gao, Y. Du, J. Liu, J. Lv, L. Yang, D. Meng, A. Hauptmann, InfAR dataset: infrared action recognition at different times, Neurocomputing 212 (2016) 36–47.

8. R. Kavi, V. Kulathumani, F. Rohit, V. Kecojevic, Multiview fusion for activity recognition using deep neural networks, J. Electron. Imaging 25 (4) (2016).

9. M. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: Proceedings of the European Conference on Computer Vision (ECCV), 2014, pp. 818–833.

10. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proceedings of the International Conference on Learning Representation (ICLR), 2014.

11. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1–9.

12. S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: Proceedings of the International Conference on Machine Learning (ICML), 2015, pp. 448–456.

13. K. He, X. Zhang, S. Ren, Jian Sun, Deep residual learning for image recogni- tion, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.

14. S. Ji, W. Xu, M. Yang, Y. Kai, 3D convolutional neural networks for human action recognition, in: Proceedings of the International Conference on Machine Learning (ICML), 2010, pp. 495–502.

15. S. Ji, W. Xu, M. Yang, K. Yu, 3D convolutional neural networks for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (1) (2013) 221–231.

16. T. Hassner, A critical review of action recognition benchmarks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 245–250.

17. M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space-time shapes, in: Proceedings of the IEEE International Conference on Computer Vision Pattern Recognition (CVPR), 2005, pp. 1395–1402.

18. M. Rodriguez, J. Ahmed, M. Shah, MACH Action, a spatio-temporal maximum average correlation height filter for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.

19. K. Soomro, A. Zamir, Action Recognition in Realistic Sports Videos, Computer Vision in Sports, Springer International Publishing, 2014.