BEHAVIOUR ANALYSIS OF MENTALLY EFFECTED PEOPLE

K. Purna Chandra Rao[1], Gandikota Gopi[2]

[1] Associate. Professor, Department of Computer Science and Engineering

kpraoakvk@gmail.com

[2] Assistant Professor , Department of Computer Science and Engineering

Venugg1989@gmail.com

[1] RISE KRISHNA SAI PRAKASAM GROUP OF INSTITUTIONS – ONGOLE

[2] RISE KRISHNA SAI GANDHI GROUP OF INSTITUTIONS - ONGOLE

ABSTRACT: Computers rely on human action recognition to comprehend human activity in visual media. There is still a lot of room for improvement in terms of accuracy since there is no background information for recognition in a single picture. This research proposes a single-image human action detection system that uses skeletal keypoints and upgraded ResNet to achieve better accuracy via a number of different means. To address the issues of a huge network and sluggish operation, we built a model based on the different images which are trained rigorously, making it more suitable for the human action recognition job while simultaneously increasing accuracy and balancing the total amount of parameters. This study evaluates ResNet, and the whole network as improvement approaches. In spite of variations in human motion, backdrop lighting, and occlusion, the findings demonstrate that the single-image human action identification system using enhanced ResNet and skeleton keypoints is capable of properly identifying human activity. .

Introduction:

Athlete action evaluation, proctoring of exams, sign language recognition based on action classification shopping mall behavior detection etc. are just a few examples of the many modern applications of human action classification (HAC), which involves computers automatically analyzing feature images and then assigning appropriate labels to the resulting images. However, owing to varied behaviors and complicated environments, action detection and categorization is a difficult subject. There has been an unceasing stream of recent HAC research based on machine learning models, which mirrors the meteoric rise in popularity of machine learning generally. To finish the action detection and classification process for 10 different classes, Liu et al. [5] used a combination of space-time features and a hyper-sphere multi-class Support Vector Machine (HS-MC SVM). By using the K-Nearest Neighbor (KNN) Algorithm, Wang et al. [6] found that, with the right choice of k, the final accuracy would be reasonably high. In order to increase classification accuracy while decreasing the negative impact on the environment, Chuan et al. presented an enhanced Random Forest model called Action Forest [7]. Ijjina and Mohan, for example, built their own convolutional neural network to automatically learn important elements of movies of human behaviors [8]. The study was based on CNNs, however. Furthermore, prior work has focused solely on the ResNet model—one of the classical and outstanding CNN models—and how to improve ResNet model's performance when dealing with HAC without incorporating any other features extraction methods. However, Yang et al. [9] aims to improve human action recognition accuracy by combining various feature extraction methods with CNN. More action categories and random backdrops warrant more in-depth research when dealing with complicated surroundings and multi-behaviors. In this research, we provide a method that uses the ResNet model to properly categorize human activity photographs into fifteen distinct groups, such as eating, using a laptop, smiling, etc. Many unknowns accompany people's acts in the actual world. The fact that there is only one class for all of these different ways people drink water—from cups to straws to just holding the water in their

hands—means that CNN models need to learn more "details" from the images they process in order to complete multi classification tasks. As the number of learning layers increases in ResNet, the gradient becomes substantially smaller, allowing the model to acquire more features at greater operating speeds [10]. Consequently, this study focuses on many major points: (1) Make sure that the training and testing sets of images for each action type include a diverse variety of styles and environments, including close-ups, far-shots, mixed-action sequences, and scenes with many people. The goal is to make sure the trained model works better in a scenario that mimics the actual world as much as possible. Using the ResNet model for this classification problem, we may improve the accuracy while minimizing the loss value by enhancing images, selecting the parameters, and increasing the number of layers.

Methodology:

The data used in this study came from the real time pictures and videos. There are two halves to the original data set: a training set with 1000 images of people in various behaviors and a test set with 700 images of people in various behaviors. There are fifteen different activity categories to which these photographs belong: "calling," "cycling," "dancing," "drinking," "eating," "fighting," "embracing," "laughing," "listening_to_music," "running," "sitting," "sleeping," "texting," and "using_laptop." The test set uses a ".csv" file that follows a one-to-one correspondence between picture names and action labels; each folder in the test set contains an average of 840 pictures. The training set uses an approach wherein each picture is first classified based on the user's subjective assessment of the picture's behavior; then, each folder in the training set contains an average of 110 pictures. All of these images are RGB photographs, and they range in size.

Each image in the training and test sets undergoes three stages of preprocessing. The ResNet50, ResNet101, and ResNet152 models imposed size restrictions on input photographs, hence it is recommended to first limit the picture size to 224×244. Next, enhance the input photographs by adjusting their brightness, contrast, and saturation levels; this will make the input pictures more accurate and less affected by blurriness. Lastly, bringing all of the images into conformity with the mean and standard deviation values that were determined from the training set augmentation. The parameter's details are shown in the table.

Table 1. The details of the parameter of preprocessing

Name of parameters	Value
size	(224,224)
brightness	0.5
contrast	0.5
saturation	0.5
mean	[0.5728639, 0.53787696, 0.50688255]
Standard deviation	[0.2453927, 0.24324809, 0.24655288]

ResNet Model

Among the top convolutional neural network (CNN) models, Deep Residual Network (ResNet) is used for a variety of applications because it attempts to avoid disappearing or explosion gradients as the network depth increases [12, 13]. Figure 2 shows the working concept of one ResNet building block. In this block, x is the final output of the previous building block, F(x) is the set of features extracted by this block, and the result of this block is represented by F(x) + x [10].

Implementation

By using GPU, the research made use of the well-known deep learning framework PyTorch, which may speed up the process of human action categorization. By keeping the initial learning rate, batch size, optimizer, loss function of train, loss function of test, and epochs constant, as well as adjusting the appropriate parameters and hyperparameters of each ResNet model, this paper aims to compare the accuracy and loss of the classification result using ResNet18, ResNet50, ResNet101, and ResNet152, respectively.

Table 2. The structure and brief information of three models mainly used in this project: ResNet18, ResNet50, ResNet101 and ResNet152

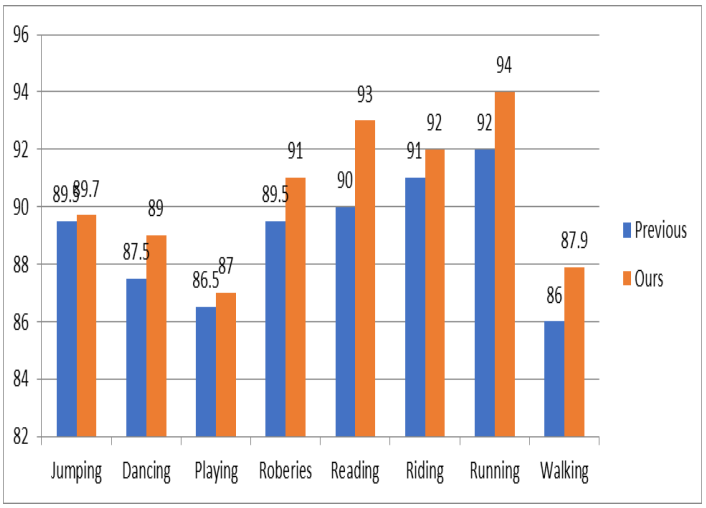
layer name	output size	18-layer	50-layer	101layer	152layer
conv1	112x112	7x7, 64, stride 2			
conv2	56x56	3x3 max pool, stride 2			
		$\begin{bmatrix}3\times3,64\\3\times3,64\end{bmatrix}\times2$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$
conv3	28x28	$\begin{bmatrix}3\times3,128\\3\times3,128\end{bmatrix}\times2$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times8$
conv4	14x14	$\begin{bmatrix}3\times3,256\\3\times3,256\end{bmatrix}\times2$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times23$	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times36$
conv5	7x7	$\begin{bmatrix}3\times3,512\\3\times3,512\end{bmatrix}\times2$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$
	1x1	average pool, 1000-d fc, softmax			
FLOPs		1.8×10^9	3.8×10^9	7.6×10^9	11.3×10^9

Results:

Using ResNet, the human action recognition test experiment was conducted in this article. Following 20 training rounds (604 iterations) on the Pascal Voc dataset, the network maintained an accuracy of about 90.5%.

After adding the points to the ResNet model, we compared its results to those of the baseline model using the keypoints detection ablation test, which adjusts for other factors. Adding the keypoints to the model employed in this research increases the classification indices compared to the prior one, as can be immediately shown.

We also checked the model's performance against other networks. For the comparative experiment, the following networks were chosen: CNN [3], Action Mask [1], CNN [4], Whole and Parts [11]. And in seven out of ten categories, the algorithm came out on top. Calling, playing an instrument, reading, jogging, photographing, using a computer, and strolling are the seven types of activities. Despite without using any extra data or gimmicks, the method presented in this work nevertheless manages to obtain a competitive result. The efficacy of the suggested approach may be shown in this way.



Comparison Graph

Conclusion:

An upgraded version of ResNet and human action recognition system is proposed to address the issue of low single-image accuracy in this area. As its core classification network, ResNet was supplemented with CPN in this approach. A multitasking architecture underpins the whole network. Based on this, we may enhance the recognition accuracy without increasing the total network parameters by modifying the backbone ResNet and branch CPN networks. The results of the experiments demonstrate that the technique outperforms previous single-image human action recognition networks.

References

1. Cippitelli, E.; Fioranelli, F.; Gambi, E.; Spinsante, S. Radar and RGB-depth sensors for fall detection: A review. *IEEE Sens. J.* 2017, *17*, 3585–3604. [Google Scholar]
2. Cai, H.; Fang, Y.; Ju, Z.; Costescu, C.; David, D.; Billing, E.; Ziemke, T.; Thill, S.; Belpaeme, T.; Vanderborght, B.; et al. Sensing-enhanced therapy system for assessing children with autism spectrum disorders: A feasibility study. *IEEE Sens. J.* 2018, *19*, 1508–1518. [Google Scholar]

3. Kong, Y.; Fu, Y. Modeling supporting regions for close human interaction recognition. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; Springer: Berlin/Heidelberg, Germany, 2014; pp. 29–44. [Google Scholar]
4. Zhang, J.; Li, W.; Ogunbona, P.O.; Wang, P.; Tang, C. RGB-D-based action recognition datasets: A survey. *Pattern Recognit.* 2016, *60*, 86–105. [Google Scholar]
5. Chen, L.; Wei, H.; Ferryman, J. A survey of human motion analysis using depth imagery. *Pattern Recognit. Lett.* 2013, *34*, 1995–2006. [Google Scholar]
6. Lun, R.; Zhao, W. A survey of applications and human motion recognition with microsoft kinect. *Int. J. Pattern Recognit. Artif. Intell.* 2015, *29*, 1555008. [Google Scholar]
7. Presti, L.L.; La Cascia, M. 3D skeleton-based human action classification: A survey. *Pattern Recognit.* 2016, *53*, 130–147. [Google Scholar]
8. Han, F.; Reily, B.; Hoff, W.; Zhang, H. Space-time representation of people based on 3D skeletal data: A review. *Comput. Vis. Image Underst.* 2017, *158*, 85–105. [Google Scholar]
9. Ye, M.; Zhang, Q.; Wang, L.; Zhu, J.; Yang, R.; Gall, J. A survey on human motion analysis from depth data. In *Time-of-Flight and Depth Imaging. Sensors, Algorithms, and Applications*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 149–187. [Google Scholar]
10. Aggarwal, J.K.; Xia, L. Human activity recognition from 3d data: A review. *Pattern Recognit. Lett.* 2014, *48*, 70–80. [Google Scholar]
11. Zhu, F.; Shao, L.; Xie, J.; Fang, Y. From handcrafted to learned representations for human action recognition: A survey. *Image Vis. Comput.* 2016, *55*, 42–52. [Google Scholar]
12. Aggarwal, J.K.; Ryoo, M.S. Human activity analysis: A review. *ACM Comput. Surv. (CSUR)* 2011, *43*, 1–43. [Google Scholar]
13. Dawn, D.D.; Shaikh, S.H. A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector. *Vis. Comput.* 2016, *32*, 289–306. [Google Scholar]
14. Zhang, Z.; Liu, S.; Liu, S.; Han, L.; Shao, Y.; Zhou, W. Human action recognition using salient region detection in complex scenes. In Proceedings of the Third International Conference on Communications, Signal Processing, and Systems, Hohhot, Inner Mongolia, China, 14–15 July 2014; Springer: Berlin/Heidelberg, Germany, 2015; pp. 565–572. [Google Scholar]

15. Nguyen, T.V.; Song, Z.; Yan, S. STAP: Spatial-temporal attention-aware pooling for action recognition. *IEEE Trans. Circuits Syst. Video Technol.* 2014, *25*, 77–86. [Google Scholar]
16. Zhang, H.B.; Lei, Q.; Zhong, B.N.; Du, J.X.; Peng, J.; Hsiao, T.C.; Chen, D.S. Multi-surface analysis for human action recognition in video. *SpringerPlus* 2016, *5*, 1–14. [Google Scholar]
17. Burghouts, G.; Schutte, K.; ten Hove, R.M.; van den Broek, S.; Baan, J.; Rajadell, O.; van Huis, J.; van Rest, J.; Hanckmann, P.; Bouma, H.; et al. Instantaneous threat detection based on a semantic representation of activities, zones and trajectories. *Signal Image Video Process.* 2014, *8*, 191–200. [Google Scholar]
18. Wang, H.; Schmid, C. Action recognition with improved trajectories. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3551–3558. [Google Scholar]
19. Oreifej, O.; Liu, Z. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 716–723. [Google Scholar]
20. Li, M.; Leung, H.; Shum, H.P. Human action recognition via skeletal and depth based feature fusion. In Proceedings of the 9th International Conference on Motion in Games, Burlingame, CA, USA, 10–12 October 2016; pp. 123–132. [Google Scholar]
21. Yang, X.; Tian, Y. Effective 3d action recognition using eigenjoints. *J. Vis. Commun. Image Represent.* 2014, *25*, 2–11. [Google Scholar]
22. Chen, C.; Liu, K.; Kehtarnavaz, N. Real-time human action recognition based on depth motion maps. *J. Real-Time Image Process.* 2016, *12*, 155–163. [Google Scholar]
23. Azure Kinect DK. Available online: https://azure.microsoft.com/en-us/products/kinect-dk/ (accessed on 6 February 2023).
24. Simonyan, K.; Zisserman, A. Two-stream convolutional networks for action recognition in videos. *Adv. Neural Inf. Process. Syst.* 2014, *27*. [Google Scholar]
25. Tran, D.; Bourdev, L.; Fergus, R.; Torresani, L.; Paluri, M. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4489–4497. [Google Scholar]

26. Liu, J.; Shahroudy, A.; Xu, D.; Wang, G. Spatio-temporal lstm with trust gates for 3d human action recognition. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 816–833. [Google Scholar]
27. Wang, P.; Li, W.; Gao, Z.; Zhang, J.; Tang, C.; Ogunbona, P.O. Action recognition from depth maps using deep convolutional neural networks. *IEEE Trans. Hum.-Mach. Syst.* 2015, *46*, 498–509. [Google Scholar]
28. Güler, R.A.; Neverova, N.; Kokkinos, I. Densepose: Dense human pose estimation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7297–7306. [Google Scholar]
29. Fang, H.S.; Xie, S.; Tai, Y.W.; Lu, C. Rmpe: Regional multi-person pose estimation. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2334–2343. [Google Scholar]
30. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299. [Google Scholar]
31. Yan, S.; Xiong, Y.; Lin, D. Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 4–6 February 2018; Volume 32. [Google Scholar]
32. Zhao, Y.; Xiong, Y.; Wang, L.; Wu, Z.; Tang, X.; Lin, D. Temporal action detection with structured segment networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2914–2923. [Google Scholar]