May - June 2020
ISSN: 0193-4120 Page No. 4209-4214

Smart Fitness Trainer System Using Computer Vision

Tejas Rao C[1], Mohammed Zainuddin[1], Syed Faraaz Ahmed[1], Shrishail M Patil[1], Priyadarshini R[2]

[1]UG Student, School of Computing & IT, REVA University, Bangalore, Karnataka
[2]Professor, School of Computing & IT, REVA University, Bangalore, Karnataka
[1]tejasraocktl@gmail.com, [1]mzainuddin28@gmail.com, [1]syedfaraazahmed31@gmail.com, [1]shrimpatil999@gmail.com, [2]priyadarshinir@reva.edu.in

Article Info
Volume 83
Page Number: 4209-4214
Publication Issue:
May - June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *12 May 2020*

Abstract

Computer Vision is a field of study that helps the computer see images and understand the content of digital images such as photographs and videos. It attempts to understand the 2D space of an image and uses that to learn the 3D nature of the object. Our goal is to develop a trainer system that visually monitors the user's movements for a particular exercise and then provides the accuracy score along as a feedback. In order to incorporate this we use simple Higher-Resolution Net (HRNet) architecture to achieve Human Pose Estimation, also known as Key point Detection and an Action Recognition model. Action recognition helps us in recognizing a human action from a video containing complete action execution. Since the action involves spatial and temporal context a 3D Convolutional Neural Network (CNN) is used. The introduced system uses UCF101 –Action Recognition dataset and other sources such as YouTube-8M for collection of the exercise dataset. 3D CNN also makes use of these datasets for training the model and for generating the class scores. The system includes the aforementioned HRNet model and the 3D – CNN for Action Recognition. Initially the system takes the video as an input from the user performing an exercise and then uses HRNet to identify the key points of that user or person and simultaneously the 3D-CNN recognizes the action class. Along with it, a keypoint scoring algorithm is used to generate scores for each keypoints or joints along with the overall accuracy of the exercise.

Keywords: *— 3D Convolutional Neural Network, HRNet, Keypoint Detection, Action Recognition, Computer Vision, Human Pose Estimation, Keypoint scoring algorithm*

1. Introduction

The ability to see and perceive information is a basic building block in any intelligent species. The same holds true for a machine as well. For the longest time, observing the human body and deducing information from it has been a challenge. Once that was achieved, the obvious next step would then be to use that to apply in daily-life applications to make life easier. Our purposed trainer system hence tries to combine all these. To achieve this we use the HRNet architecture using the Human Pose Estimation model also known as Keypoint Detection. Pose Estimation is where we detect the position and orientation of an object. This means detecting keypoint locations that classify the object. Human Pose Estimation attempts to find the orientation and configuration of human body parts like feet, shoulders, knees, etc. Athletic activities like basic warm up exercises are a necessity to maintain physical fitness. However not everybody can afford to visit a gym or have a physical trainer, hence our system proves to be a handy tool that can be used at the convenience in their homes.

A 3D CNN is similar to a 2D except that the 3D convolutions are performed to compute the features from both spatial and temporal dimensions rather than the 2D convolutions that compute features from the spatial dimensions only. Action recognition problem involves the identification of various actions from video clips (a sequence of 2D frames) where the action may or may not be performed throughout the whole duration of the video. For accomplishing this task 3D CNN is employed in sync with UCF101 dataset.UCF101 is an action recognition dataset of practical action videos, collected from various

May - June 2020
ISSN: 0193-4120 Page No. 4209-4214

videos on YouTube, having 101 action categories. This data set is an add-on of UCF50 data set which has 50 action categories. The UCF101 dataset provides 8 different exercises classes which are bench press, bodyweight squats, handstand pushups, jumping jacks, lunges, pullups, pushups and wall pushups. Each class contains over 100 clips. These clips are used to train 3D CNN through transfer learning.

The keypoints dataset are generated from the clips of different exercise classes. These generated keypoints are required in keypoint scoring algorithm. The keypoint scoring algorithm performs the main function of scoring the accuracy of the exercise. This algorithm takes each frame in the video input of the user and provides an accuracy score for each different major joints. These major joints are chosen based on the exercise.

The remaining of the paper is sorted out as pursues: Section II provides the review of work done in building the fitness trainer system utilizing various Computer Vision techniques. Section III highlights the theoretical background of the proposed architecture followed by results and discussion in Section IV. The conclusion and future work is talked about in Section V. End and future work bearings are referenced in Section VI.

2. Literature Survey

Zhe Cao, Tomas Simon, Shih-En Wei and Yaser Sheikh[1]introduce 2D pose estimation model that detects the poses and keypoints of multiple persons in a single image. This is achieved through Part Affinity Fields, which is a non-parametric representation to associate body parts with individuals. This uses a greedy bottom up approach that maintains high accuracy while achieving real-time performance. The architecture of this estimation model is designed in such a way to mutually learn the part locations and their association across two branches, which undergo same sequential prediction process.

The learning of High-resolution representation has a crucial part in many vision problems, e.g., semantic segmentation and pose estimation. The high-resolution network (HRNet) [3], maintains high-resolution representations throughout the whole procedure. Ke Sun et al. [2] have conducted an evaluation on high-resolution representations by initiating a direct but functional modification and apply it to a wide range of vision tasks. By accumulating the (upsampled) representations from all the parallel convolutions they augment the high-resolution representation. This simple modification results in stronger representations, indicated by superior results.

In this paper, Ke Sun et al. [3] use high-resolution representations in human pose estimation problem. Most of the previously existing methods used high-to-low resolution network to recover high-resolution representations from low-resolution representations. This model maintains high-resolution representations through the whole process. The network starts of as a high-resolution subnetwork as the first phase, then gradually adds high-to-low resolution subnetworks sequentially to form more phases, and connect the multi-resolution subnetworks in parallel. They repeat multi-scale fusions such that each of the high-to-low resolution representations receives information from other parallel representations over and over, leading to rich high-resolution representations. As a result, the predicted key-point heat map is more accurate and spatially more precise.

Sen Qiao, Yilin Wang, Jian Li, Qiao, S., Wang, Y., & Li, J[4] presents a real-time 2D human gesture grading system based on OpenPose, a library for real-time multi-person keypoint detection. After encapsulating 2D positions of a person's joints and skeleton wireframe of the body, the system calculates the equation of motion trajectory for every joint. Similarity metric gives distance between motion trajectories of standard and real-time videos. A modifiable scoring formula is used for simulating the gesture grading scenario.

B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang and L. Zhan[5] introduce Higher-Resolution Network which follows bottom-up approach for multi-person human pose estimation and it is more oriented towards increasing heat map prediction accuracy. The base model for this higher net is the simple HRNet [3]. The main difference when compared to the existing HRNet is that it has higher-resolution feature maps which becomes the reason to detect small and medium persons more accurately. Then, it builds multi-level features with high-quality which is used to for multi-scale pose prediction. While using higher-resolutions there computation requirements is also increased. This model is more accurate than all the currently existing bottom-up methods.

K. Soomro, A. R. Zamir and M. Shah [6] introduce a dataset known as UCF101 which contains a large amount of clips for various kinds of human actions. It consists of more than 13 thousand clips of 101 different action classes. The datasets are more practical since it is uploaded by the users itself where clips are dynamic and has cluttered background. Since it contains a lot of different action classes and large amount of clips with poor lighting, cluttered background and dynamic camera motion, it also came to known as one of the most challenging dataset for actions. The videos are mostly gathered from sources such as YouTube and movies.

K. Hara, H. Kataoka and Y. Satoh [7] propose a study to learn whether the existing video datasets have sufficient data to train deep CNNs which also contains spatio-temporal data. The addition of spatio-temporal data adds one more dimension to the 2D kernels and making it a 3D kernel, hence a 3D CNN. Based on the history of 3D CNNs, they are of the opinion that these nets have made a significant presence in the field of action recognition, where the CNNs are mostly made of shallow 3D architectures. But they explore the architectures from shallow to very deep ones on existing

May - June 2020
ISSN: 0193-4120 Page No. 4209-4214

video datasets. And they finally concluded that Kinetics dataset has enough information to train 3D CNNs with up to 152 layers and are more accurate when compared to existing 2D architectures. But datasets such as UCF-101 [6], HMDB-51 and ActivityNet overfits when trained on ResNet-18 which has lesser layers.

K. Hara, H. Kataoka and Y. Satoh [8] propose the study of using 3D Residual Networks (ResNets) to learn spatio-temporal features for action recognition. 3D CNNs uses spatio-temporal 3D kernels and can directly extract spatio-temporal features from clips for action recognition. 3D CNNs usually overfits due to the reason that it has huge number of parameters, but this is overcame by using a huge video datasets. They mention the procedure to train a 3D CNNs based on ResNets and trained results for 3D ResNets on Kinetics dataset which contains a large number video clips. The results show that 3D ResNets has better performance in spite of the network having a large number of parameters. 3D ResNets showed a better result when compared to existing shallow architecture based networks, such as C3D.

3. Methodology

The primary objective of our project is to build a Smart Fitness Trainer system, which can be used by anyone at their convenience to practice various fitness exercises and then obtain an accuracy score for the exercise performed. The input for this project is the video of the user performing the exercise, which should at least have a minimum number of frames involving the complete execution of at least a single rep of exercise. The output is the final output score given along with the feedback. The proposed methodology is given in Fig 1. The modules used in the methodology includes- Input, Pose Estimation Model, 3D CNN Model and the datasets are further divided into the ones used to train the HRNet and the ones used to train the 3D CNN.After that follows the last module - Keypoint Scoring algorithm.

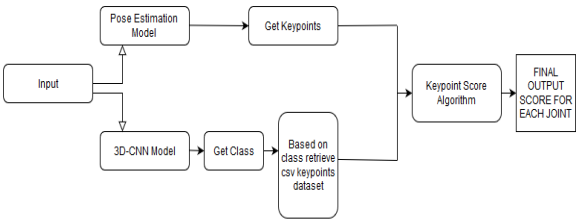

Figure 1: Workflow of the project

A. Input

The input is a raw video that we receive from the user. As mentioned earlier the video should at least have aminimum number of frames involving the complete execution of at least a single rep of exercise. The reason is that the model can provide a more accurate score once the user completes a single rep. The input is then sent to two different models i.e., Pose Estimation Model and 3D CNN Model.

B. Pose Estimation Model

This is a Human Pose Estimation model that initially receives the input from the user. This model is primarily used to detect keypoints. The Human Pose Estimation can be described as the localization of human joints or keypoints such as elbows, wrists, etc. The pose estimation model has wide applications. The best model that utilizes the Pose Estimation technique is the Deep High-Resolution Representational Learning for Human Pose Estimation using HRNet [3]. In terms of accuracy, this model has outperformed every other model in Keypoint Detection. The rationale behind this is that majority of the previously existing methods form low-resolution representations recover high-resolution representations produced by a high-to-low resolution network. This model maintains high-resolution representations throughout the entire process.

The datasets that are used to train this model mostly consist of images. The model is further tested for accuracy on two benchmark results i.e., the COCO keypoint detection dataset and the MPII Human Pose dataset. The average recall of the HRNet model when compared with COCO dataset is a high of 82.0%, which is the highest among all other models. The keypoint accuracy when compared with MPII test set has 92.3% accuracy.

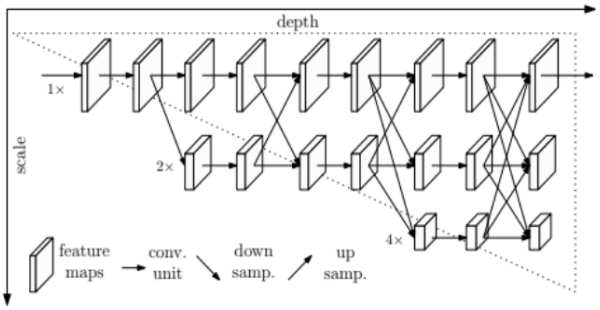

Figure 2: HRNet architecture model

Figure 3: 2D Pose Estimation

C. 3D-CNN Model

This is the other model where the input video is received. This model consists of a 3D Convolutional Neural

Network. A 3D CNN is generally used for action recognition. The model is primarily similar to 2D CNN but has an extra dimension to it in the form of time-series data or temporal data, which introduces a 3D kernel. We trained 3D CNN model using transfer learning. Transfer learning [9] is a machine learning concept where a pre-trained model is used for another task to jump-start the developmental process for a new problem. We used an existing 3D CNN trained on UCF101 dataset that contains 101 action class categories. We re-train the 3D-CNN on an exercise dataset that contains eight different action categories. The exercise dataset is gathered sources such as, UCF101 and YouTube-8M.The output provided by this model is an exercise class label. We were able to achieve an accuracy of 89.13%.

Layer Name	Architecture
	34-layer
conv1	7 x 7 x 7, 64, stride 1 (T), 2 (XY)
conv2_x	3 x 3 x 3 max pool, stride 2 $\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \end{bmatrix}$ x 3
conv3_x	$\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix}$ x 4
conv4_x	$\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix}$ x 6
conv5_x	$\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix}$ x 3
Average pool, 8-d fc, softmax	

Figure 4: 3D CNN Network Architecture.

Residual blocks are shown in brackets. Each convolutional layer is followed by batch normalization [10] and ReLU [11]. Downsampling is performed by conv3_1, conv4_1, conv5_1 with a stride of 2. The dimension of last fully-connected layer is set for the Exercise dataset (8 categories).

D. Keypoint Scoring Algorithm

The video dataset gathered for 3D CNN is used to create keypoints dataset for all exercise categories through HRNet. Keypoints dataset contains co-ordinates of all joints for each different exercise and for each video frame. The keypoint dataset is pre-processed to remove outliers and normalize the values. By applying density based clustering algorithm to create a cluster for each different keypoint and for each exercise. The result of pre-processing provides a set of well-formed clusters for each different keypoint per exercise.

Based on the exercise class label provided by 3D CNN, the keypoint clusters are chosen for the scoring algorithm. The Keypoint Scoring algorithm provides accuracy scores for each keypoint or joint for each frame. This algorithm takes keypoints predicted from the HRNet and the keypoint clusters selected based on exercise class as an input. It compares each input keypoint to the existing keypoint cluster data and provides a score based on the closeness to that particular cluster. The closeness is computed by finding the Euclidean distance between the centroid of the cluster and the keypoint. For further frames, the score is the average of the current score and the score of previous frames.

4. Implementation

The steps involved in the Smart Fitness Trainer are as follows:

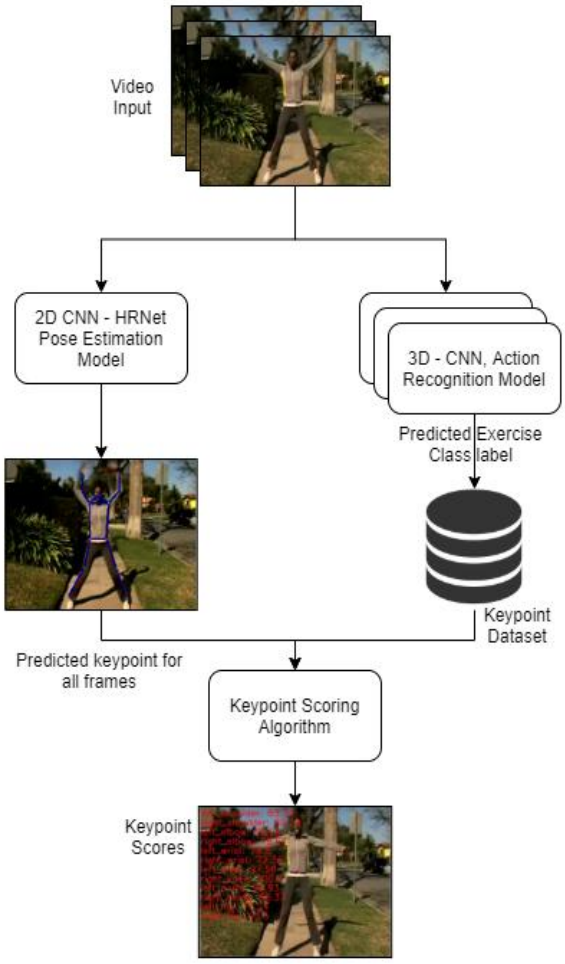

Figure 5: Dataflow diagram

Step 1: Input

- Receive an input video ***V*** from the user that involves the complete execution of at least a single rep of exercise.

May - June 2020
ISSN: 0193-4120 Page No. 4209-4214

- Forward the video to HR Pose estimation model and 3D-CNN action recognition model.

Step 2a: Identify Keypoints

- Initialize HRNet with pre-trained weights W_{hr}.
- Identify and store the keypoints K_i for each input frame V_i by passing it to the HRNet model.

Step 2b: Exercise Classification

- Initialize 3D-CNN model with pre-trained weights W_{3d}.
- Forward the entire input video V to the 3D-CNN model to identify the exercise class label C_j.

Step 3: Generate Keypoint Scores

- Based on the exercise class label C_j obtained in Step 2b retrieve the keypoints dataset D_j.
- Initialize the Keypoint scoring algorithm with dataset D_j.
- For each keypoint K_i generated in Step 2a, apply the Keypoint scoring algorithm to generate keypoint scores S_i.
- Return the keypoint scores S_i as the result.

5. Results and Discussion

To evaluate the performance of the individual components of the system such as HRNet and 3D CNN. We used 80% of the dataset to train and 20% for testing. We were able to achieve an accuracy of 86.23% for HRNet including both COCO and MPII dataset. We were able to achieve an accuracy of 89.13% for 3D CNN based on UCF101 and YouTube-8M dataset. To evaluate the performance of the Keypoint Scoring algorithm on scores or closeness, we tested each individual exercise with valid and invalid exercises. For valid exercises, the keypoint scores on average were above 60% and for invalid exercises, the keypoint scores on average were below 25%.Based on the results and dataset used to achieve the keypoint scoring results, we found that angle and position of the person needs to be align with the existing dataset. Hence, the input is one of the most important part to focus on to provide an accurate score.

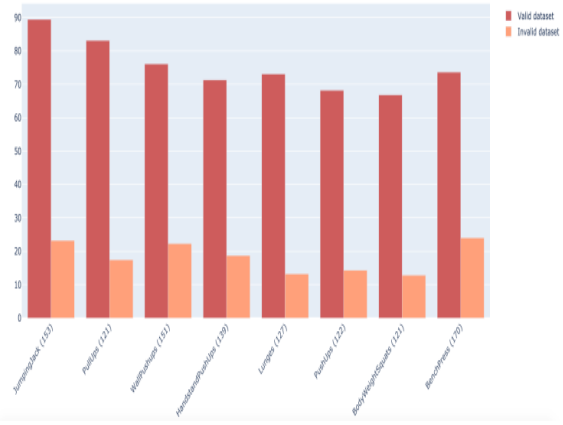

Figure 6: Accuracy results for Keypoint scoring algorithm

Figure 7: Keypoint Scoring for Jumping Jacks Exercise

6. Conclusion and Future Work

The smart fitness trainer system can be used for some basic exercises. Currently the system only supports for single-person scoring evaluation. This can be further extended for multi-persons evaluation. We have studied the Pose Estimation Model and learnt many features of keypoints, which can be used for many other use cases as well. This model even though does not require high computational capabilities, having moderately high capabilities will certainly improve the speed of the whole process. Based on the application, its functionalities can be increased as well. The scoring system can be improved based on inputs in different environments such as, different angles, exposure, etc.

References

[1] Z. Cao, T. Simon, S.-E. Wei and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," in CVPR 2017, 2017.

[2] K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu and J. Wang, "High-Resolution Representations for Labeling Pixels and Regions," arXiv, 2019.

[3] K. Sun, B. Xiao, D. Liu and J. Wang, "“Deep High-Resolution Representation Learning for Human Pose Estimation," in CVPR 2019, 2019.

[4] S. Qiao, Y. Wang, J. Li, Q. S., W. Y. and L. J., "Real-Time Human Gesture Grading Based on OpenPose," in 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI) 2017, 2017.

[5] A. Bowen, X. Bin, W. Jingdong, S. Honghui, H. S. Thomas and Z. Lei, "Bottom-Up Higher-Resolution Networks for Multi-Person Pose Estimation," arXiv, 2019.

[6] K. Soomro, A. R. Zamir and M. Shah, "UCF101: A Dataset of 101 Human Action Classes From Videos in The Wild," in CRCV-TR-12-01, 2012.

May - June 2020
ISSN: 0193-4120 Page No. 4209-4214

[7] K. Hara, H. Kataoka and Y. Satoh, "Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?”," in CVPR June 2018., 2018.

[8] K. Hara, H. Kataoka and Y. Satoh, "Learning Spatio-Temporal Features with 3D Residual Networks for Action Recognition," in ICCV October 2017., 2017.

[9] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu and H. Zhu, "A Comprehensive Survey on Transfer Learning," arXiv, 2019.

[10] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," arXiv, 2015.

[11] V. Nair and G. E. Hinton , "Rectified linear units improve restricted boltzmann machines.," in International Conference on Machine Learning , 2010., 2010.