Digital Twin Techniques in Recognition of Human Action Using the Fusion of Convolutional Neural Network

PRESENTED BY: ABIR EL

2023-03-06

CONTENTS

Introduction

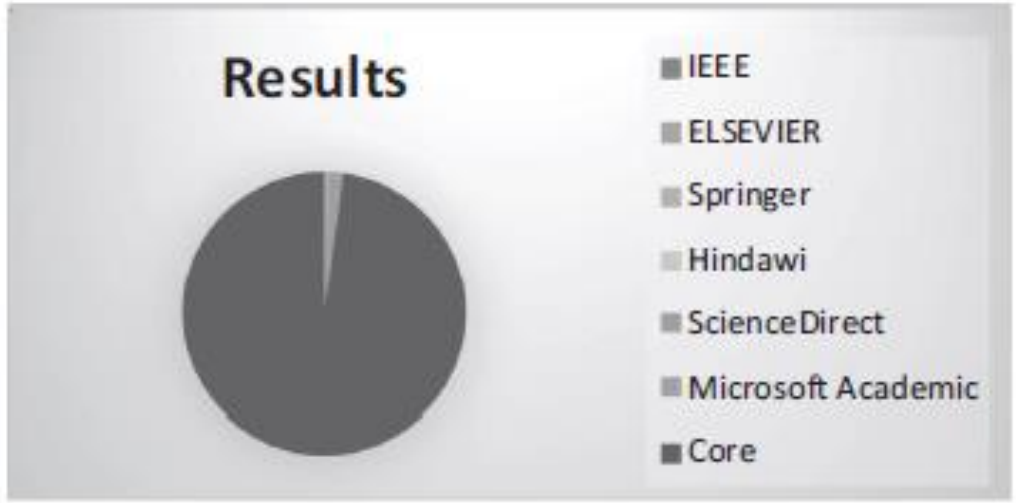
- The digital twin has increased the curiosity of many researchers recently.
- It produces a copy of the physical world into the virtual world that works the same way as the original product performs in real-time. If there is an adjustment in the virtual room, the same adjustment will implicate the real project in real-time or vice-versa.
- The digital twin technology improves ongoing operation, tests the new product, increases efficiency, and saves time and money. Digital twin technology has much application in the real world as in industry, healthcare, agriculture, aerospace, etc.
- **Recognition of action** plays a crucial role in receiving changes in the physical space, to know the actions a human performs.

Introduction

- In a real-world setting, action recognition has been an active field of research for decades as shown in Table 8.1 and Figure 8.1.

TABLE 8.1
Tabular Representation of Human Action Recognition on a Different Platform

Research Journals	Results
IEEE	4,727
ELSEVIER	11,815
Springer	47,942
Hindawi	10,000
ScienceDirect	**29,6163**
Microsoft Academic	2,106
Core	16,736,177



Introduction

- It addresses many subjects, such as video human identification, human pose estimation, human tracking, and time series interpretation and cognition.
- The current study focuses on realistic datasets gathered from movies, web videos, television programs, etc. Action identification has various benefits, such as monitoring human emotions, detecting appropriate activity, and many more.
- But in the domain of computer vision and machine learning, obtaining appropriate facts is difficult. For the best result, one must know about the behavior and effects of the human surrounding.

HISTORY OF HUMAN ACTION RECOGNITION AND DIGITAL TWIN

- Early studies were motivated by human representation in the arts by Da Vinci. According to him, human actions are characterized by how different body parts move relative to every other.
- Familiarity with the anatomy of nerves, bones, muscles, and sinews helps to understand various motions with their levels of strength.
- Gunnar Johanson used image order for programmed human motion analysis. Johansson showed that human observers can recognize biological motion patterns with moving light displays (MLD), attached to the body parts, even when presented with few moving dots.
- Later, image segmentation based on skin-color and shape analysis and the invariant moments are combined. The features are extracted and used for an input vector to a radial basis function network (RBFN).
- Real-time analysis of video stream is used to detect human motion and extract its boundaries. A human activity analysis was done in a complex environment.

HISTORY OF HUMAN ACTION RECOGNITION AND DIGITAL TWIN

- Recognizing humans' action in 3-D through skeletons was proposed. A multimodal dataset was used for human action recognition that utilized a depth camera and a wearable inertial sensor along with representation learning of temporal dynamics for skeleton-based action recognition.

- Task-based control and human activity recognition for human-robot collaboration was conducted. Deep CNN-based data-driven recognition of cricket batting shots was also conducted.

- WiAct is a passive wi-fi based human activity recognition system that works on various neural networks. A CSI-based system using Wi-fi for human activity recognition was conducted to analyze common human behaviors.

- Still, many other types of researches are going on to increase the performance of human action recognition. Figure 8.2 shows the changes shown in the field of recognizing human action.

HISTORY OF HUMAN ACTION RECOGNITION AND DIGITAL TWIN

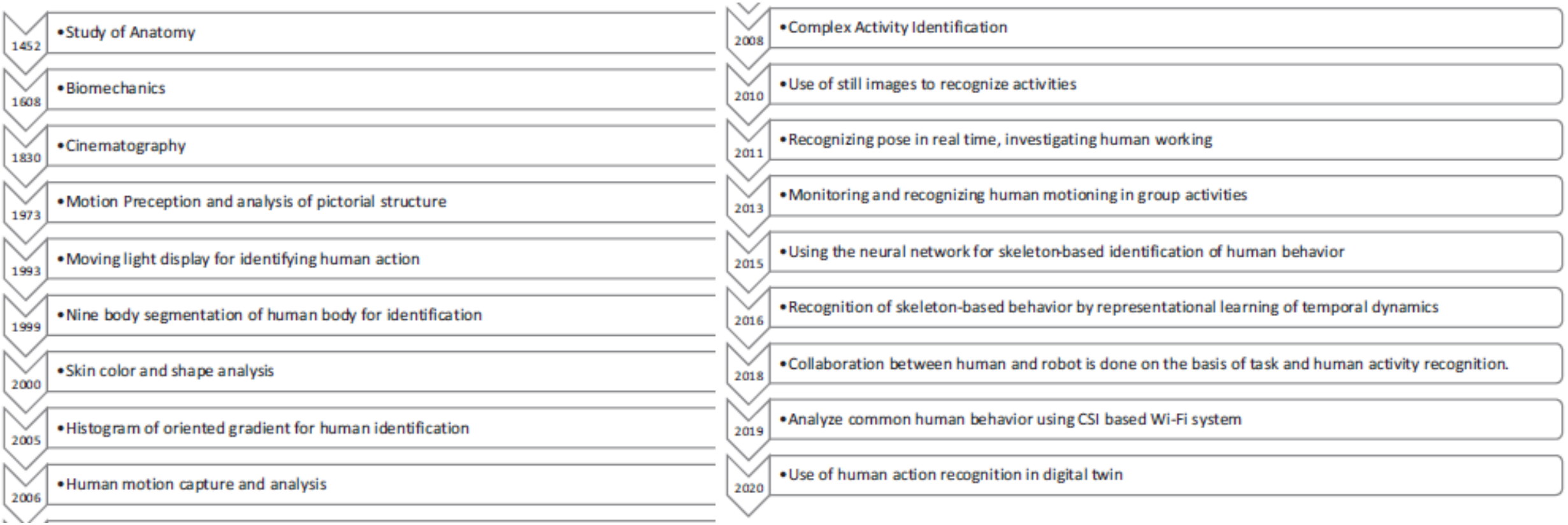


FIGURE 8.2 Evolution of human action recognition.

RELATED WORK

- The human action recognition technique flourishes in fields like video surveillance system, robotics for human behavior characteristics, multiple activity recognition systems, and recently in the digital twin environment. Different researchers have studied various parts to improve human action recognition and have made remarkable progress.

- Shotton et al. [27], proposed a method for quickly and accurately estimating 3D positions of skeletal joints using a single depth image. Here, skeleton-based action recognition approaches can be grouped into two main categories: **joint-based** and **body part-based**.

- Vemulapalli et al. [13], inspired by relative geometry between various body com-ponents, provided a more meaningful description than their absolute locations. They used the MSR-Action dataset and UTkinect-Action dataset for their experiment.

- Mahshid Majd and Reza Safabakhsh proposed to change LSTM units to cover all three types of information needed to categorize action classes in a video. Convolution and correlation operators are used to form a part, capable of extracting spatial and motion features and their temporal dependencies. They tested the proposed unit called CCLSTM, in a deep architecture with convolutional layers on UCF101 and HMDB51 datasets.

RELATED WORK

- Xikun Zhang et al. [19] noted that the movement of limbs is important to understand behavior. Given this finding, for skeleton-based action iden-tification, they explored the dynamics of human limbs.

- In a graph for the human skeleton, they represented an edge by combining its spatial coordinates to encode the coordination between different wings and temporal-adjacent margin to achieve stable movements of motion. By adding various mutual intermediate layers to com-bine graph node and edge CNNs, they further created t wo-hybrid networks.

- Graph edge convolution and hybrid network integration have been tested on Kinetics and NTURGB+D data sets for edge convolution and traditional node convolution.

- To prove the principle of adaptive route planning through a dynamic protective cover using a digital twin approach, Klaus Dröder et al. [33] suggested an experimental simulation framework for human-robot interaction.

RELATED WORK

TABLE 8.2
Experimental Result Acquired Using Different Models

Paper	Highlight	Method	Dataset	Outcomes and Future Work
Shotton et al. (2013) [27]	From single depth image prediction of human pose, it is done with the 3D position of human body joints. The whole skeleton parts are divided for accuracy of the competitive test set.	Randomized decision forests	mocap data	The findings also provide a high correlation between real and given data between classification of the intermediate and the final joint.
Vemulapalli et al. (2014) [13]	Using rotations and translations that use Lie group by explicit modeling, and 3D skeleton geometric relationships are carried out between different body sections.	3D skeleton model, Lie group model	MSR-Action3D, Kinect-Action and Florence3D-Action	Recognition of skeleton-based human behavior strategies in all three datasets, with an accuracy of about 90%
Du et al. (2015) [28]	The human skeleton is divided into five sections and then fed into five sub-networks.	Hierarchical bidirectional and unidirectional RNN	MSR Action 3D, Berkeley MHAD, and HDM05	It was not possible to separate the planned work from just the skeleton joints.
Du et al. (2016) [15]	In order to identify the behavior, hierarchical RNN is used to derive the representation of temporal dynamic of skeleton orders.	RNN and LSTM	ChaLearn, HDM05, Berkeley MHAD and MSR-Action3D	The appearance and scene information was lacking in the proposed work, which discriminates significantly in consideration of action.
Majd and Safabakhsh, (2019) [29]	Using convolution and correlation operators that are able to extract spatial and motion features and their temporal dependencies, the new unit was formed.	Correlational Convolutional LSTM	UCF101 and HMDB51	From correlation details, it was intended to analyze different unit architectures. For good performance, the network architecture could be changed.
Yan et al. (2019) [20]	Recognizing human activity using CSI-based Wi-fi to analyze common human behaviors.	Channel State Information (CSI), Adaptive Activity Cutting Algorithm (AACA) and Extreme Learning Machine (ELM)	-	In the proposed work doppler shift, correlation values are used in WiAct for ELM. Results of the ELM show high accuracy of 94.20%.

(Continued)

RELATED WORK

TABLE 8.2 (*Continued*)
Experimental Result Acquired Using Different Models

Paper	Highlight	Method	Dataset	Outcomes and Future Work
Zhang et al. (2019) [19]	By using a graph edge convolution network for action recognition, the proposed work represents each edge by integrating its neighboring edges.	Graph Edge Convolutional Neural Networks, Order-Level Hybrid Model and Body-Part-Level Hybrid Model	NTU-RGB+D Kinetics	It captures the correlation and dependencies between human limbs.
Nadeem et al. (2020) [35]	Using ANN to monitor and identify human behavior based on recognition of body parts.	Linear discriminant and ANN	Weizmann Human action dataset and KTH-dataset	As these acts are special and multidimensional functions, the confusion matrix for KTH-dataset offers 100% accuracy for running with the ball, dribbling, and standing. In the meanwhile, take a pass, shoot, run, and pass. The mean accuracy of identification is 87.57%.
Zhao and Jin (2020) [34]	The proposed work uses Bi-LSTM model of p-non-local and Fusion Key Less Attention to achieve human behavior.	Bi-LSTM No-local and CNN model	HMDB51	The p-non-local block reduces the computational complexity of long-distance dependencies, ensuring efficiency.
Xuemin et al. (2020) [36]	The built-in twin-based digital assembly, commissioning total factor information model for aerospace style, high-precision, electro-hydraulic, servo valve.	Assembly Commission method, Pareto optimal method, and Digital twin model	-	Assembly accuracy of the military product was lower than the micron level and there is no deep discussion about upstream and downstream.

ARCHITECTURAL FRAMEWORK OF HUMAN ACTION RECOGNITION IN DIGITAL TWIN TECHNOLOGY

- The digital twin technology allows the physical and virtual world to communicate, as shown in Figure 8.3.

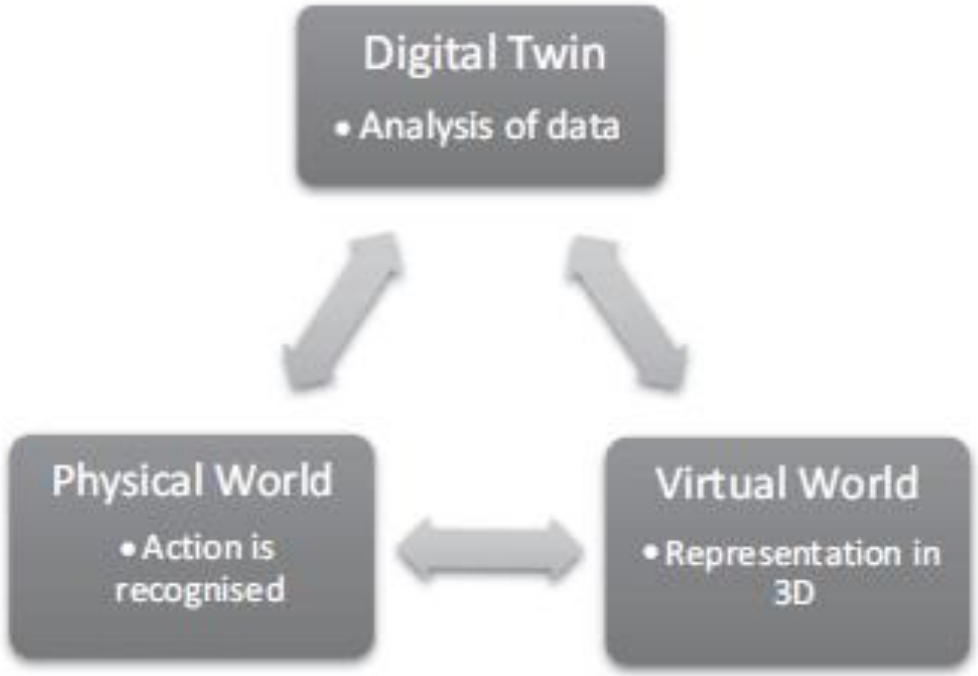


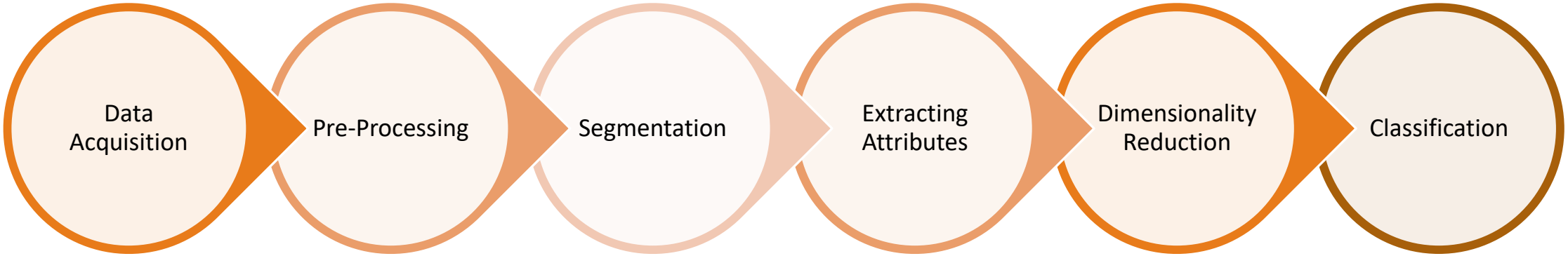
FIGURE 8.3 Digital twin model.

ARCHITECTURAL FRAMEWORK OF HUMAN ACTION RECOGNITION IN DIGITAL TWIN TECHNOLOGY

- The concept behind it is that it provides a platform where the non-living objects can show a living behavior.
- For example, if we build a twin of a city, then the working will be the same as at the original site.
- It will monitor by taking all the necessary data and analyze its future necessities or capabilities.
- If any kind of problem is there, it will be able to predict for the coming future.
- The digital twin layer will inform beforehand and will also suggest the best way to solve it.
- Right now, it asks for suggestions from humans but in the coming future, it might be able to make its decision without human implication. The first step for this goal is to be able to find the activity performed in its surrounding.

ARCHITECTURAL FRAMEWORK OF HUMAN ACTION RECOGNITION IN DIGITAL TWIN TECHNOLOGY

- Human Action Recognition is categorized into various steps. Figure 8.4 shows the architecture of human action recognition.



Data Acquisition

- Recognition of human behavior (Figure 8.5) may be sensor-based or vision-based. Knowledge of human behavior from a data perspective is limited to developments that involve RGB data, depth data, sensor data, or skeleton data

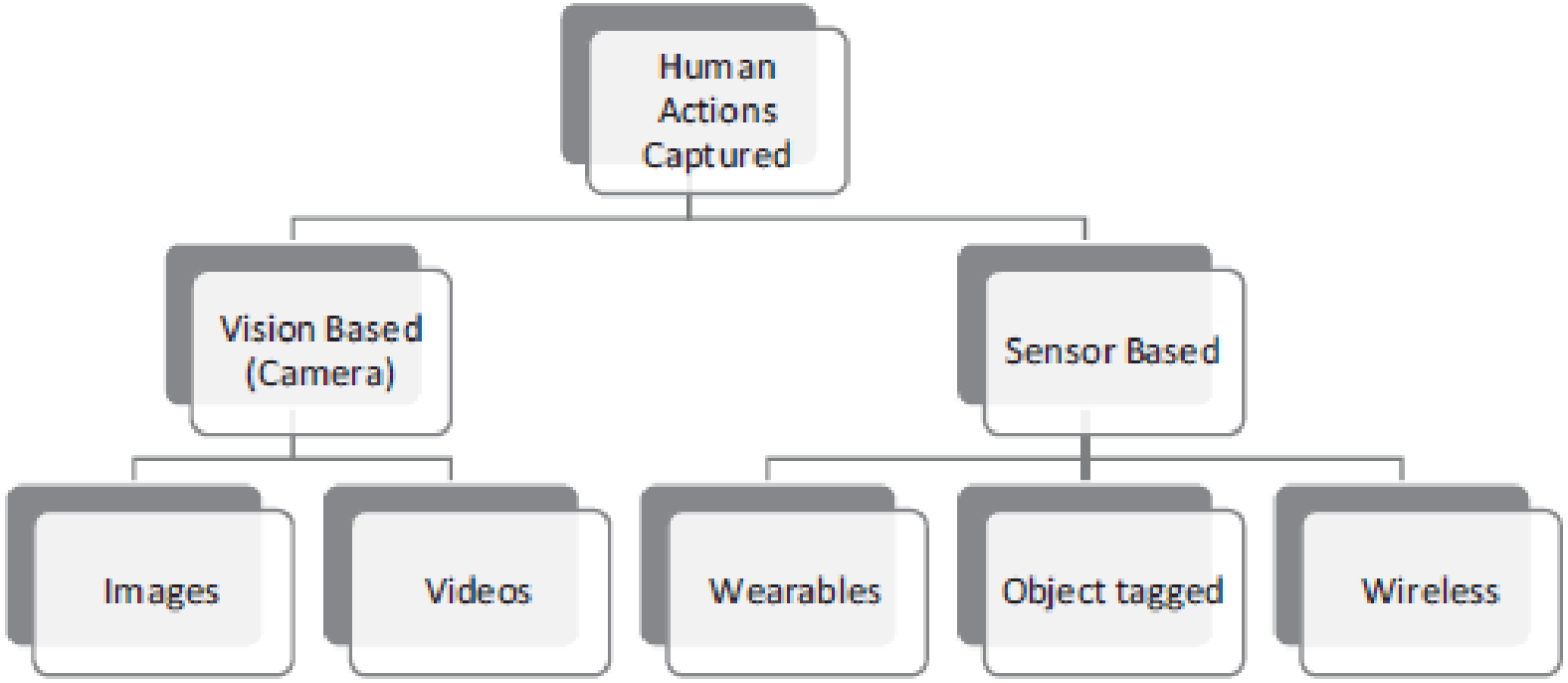


FIGURE 8.5 Data acquisition.

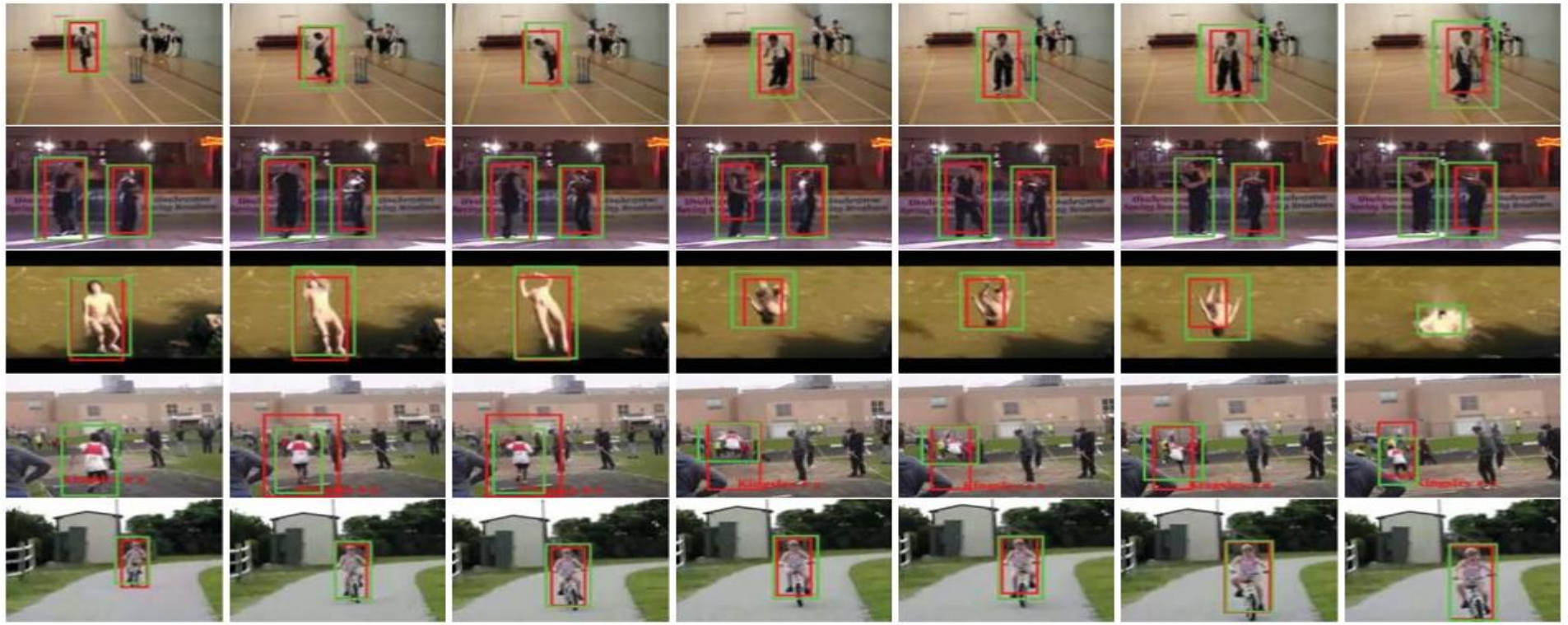
Pre-processing

- Data that we gather need to be pre-processed before further analysis. As data can be incomplete, noisy, inconsistent, and may not have the quality which can be useful, the data mining process is required which performs the following tasks as below:

  - ✓ **Data cleaning**: It is the process of removing or modifying incorrect, corrupt, duplicate, or incomplete data.
  - ✓ **Data integration**: It means combining information from various sources into a unified view. It efficiently manages data and makes it available to those who need it.
  - ✓ **Data transformation**: It converts data from one format to another.
  - ✓ **Data reduction**: It reduces the amount of capacity required to store data and increases storage efficiency at reduced cost.
  - ✓ **Data discretization**: It converts the continuous data that attributes values to a discrete counterpart.

Segmentation

- After the raw data is being pre-processed, it is required to be **segmented**. Image segmentation provides a powerful semantic description of video imagery, essential in image understanding and efficient manipulation of image data. In particular, segmentation based on image motion defines regions undergoing alike motion, which allows the image coding system to effectively show video order. Figure 8.6 shows data segmentation from the data from a different location.



Extracting attribute

- Attribute extracting is to automatically extract features from signals or images by creating new features from the existing ones. The dimensions by which an initial collection of raw data is reduced to more manageable groups for processing are reduced in this process. Figure 8.7 shows data extraction, where each part is extracted from the data used.

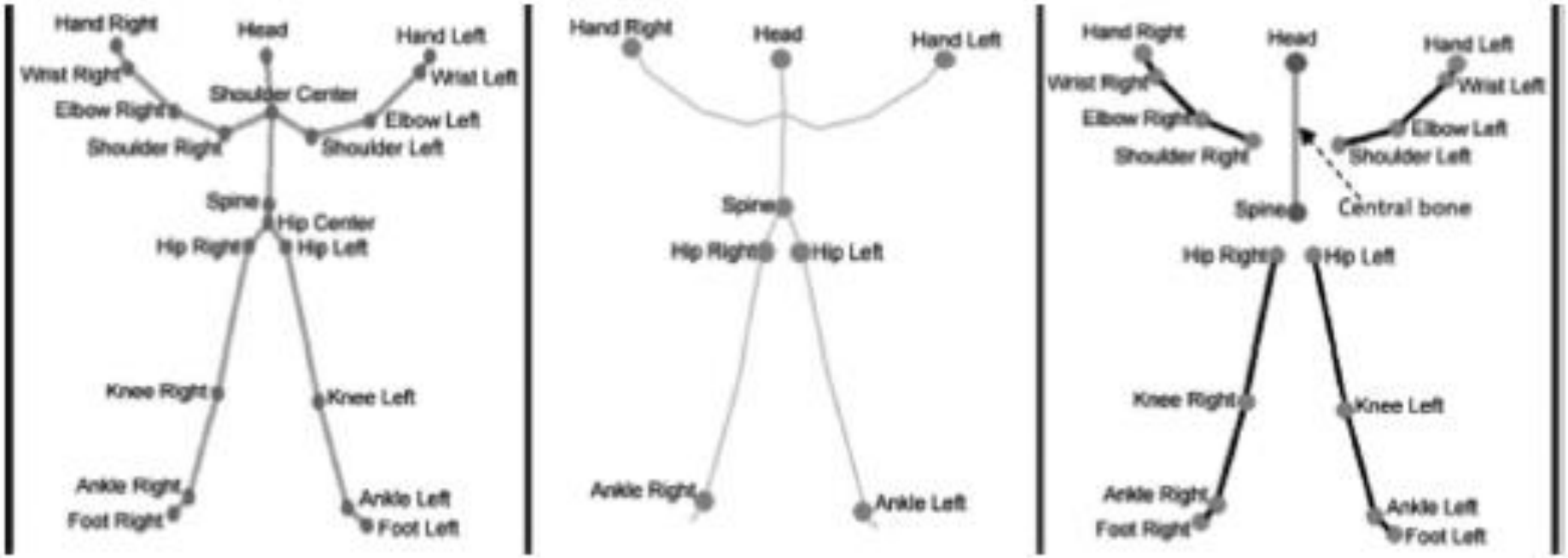


FIGURE 8.7 Data extraction [39].

Dimensionality reduction

- Dimensionality reduction refers to techniques for reducing the number of input variables in training data.
- When dealing with high dimensional data, it is often useful to reduce dimensionality by projecting the data to a lower-dimensional subspace which captures the essence of data.
- Avoiding overfitting is a major motivation for performing dimensionality reduction.

Classification

- Classification is a supervised approach (as shown in Figure 8.8) wherein the computer program learns from the data it receives and makes discoveries or classifications.

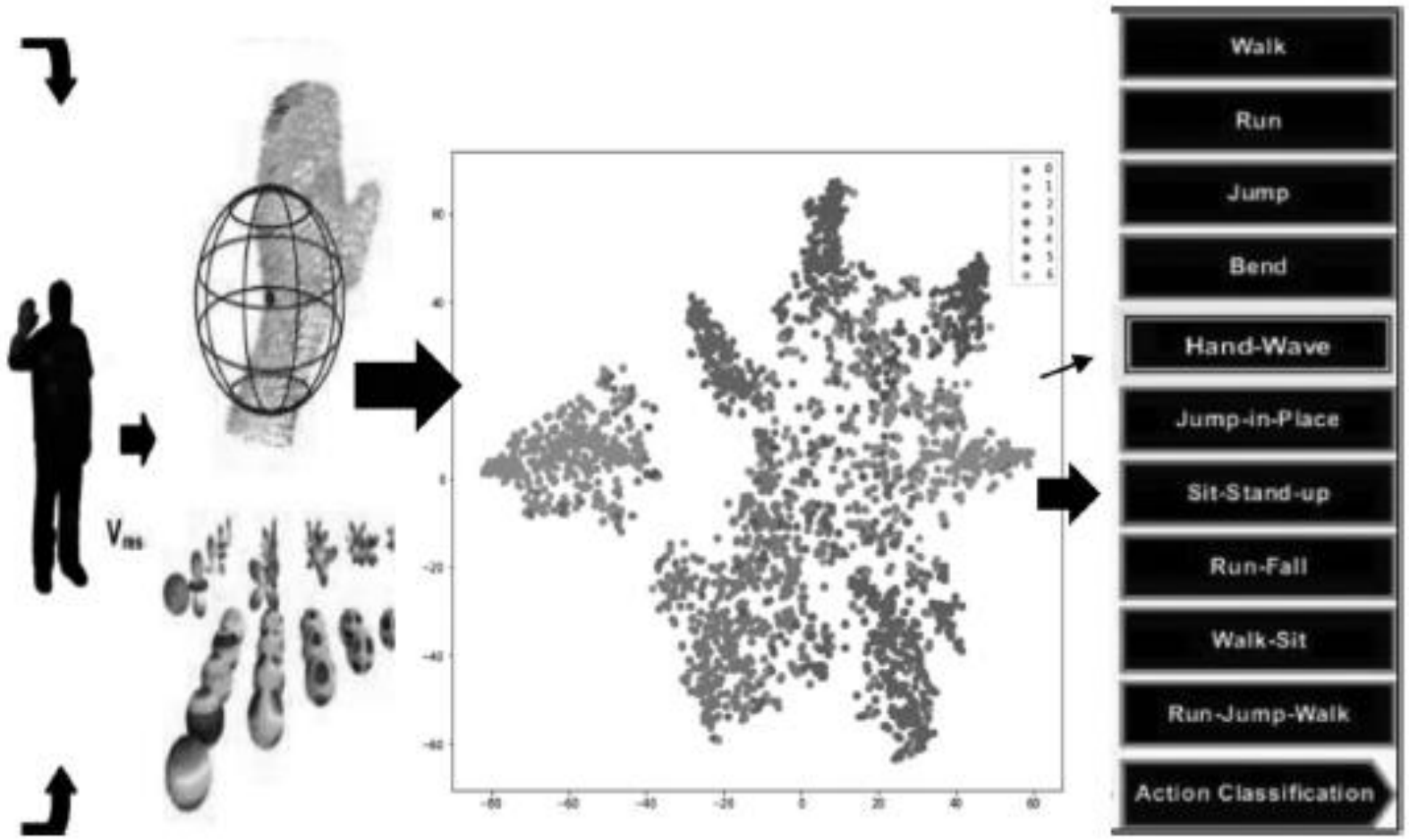


FIGURE 8.8 Data classification.

CHALLENGES FACED IN HUMAN ACTION RECOGNITION IN DIGITAL TWIN ENVIRONMENT

- The identification of actions taken by humans is increasing and is applied in several companies. Although it has positive implications, there are some drawbacks also, as given below:

  ✓ **Background jumble:**
  - The background jumble suggests that there are human beings present in the picture and it may be difficult for anyone to concentrate on a single person. Recognition of action is itself a difficult issue and the inclusion of background jumble makes the issue more difficult as it would be difficult to distinguish the person in the video from the freely moving individuals. The main area of the picture is obscured by the shadow.

  ✓ **Partial obstruction**:
  - Obstruction of a local region of the body with objects such as sunglasses, scarves, hands, watches, etc. is generally called partial occlusions. The obstruction generally has to be less than 50% to be scrutinized as a partial occlusion.

CHALLENGES FACED IN HUMAN ACTION RECOGNITION IN DIGITAL TWIN ENVIRONMENT

- ✓ **Scale change** :
  - One of the major issues in dissipated environments is that of scale; the action of a human captured at large distances is considerably harder to recognize than the same human's action at small distances. This problem is very common in image classification.

- ✓ **Perspective**:
  - Perspective variation occurs when the same image is viewed from different angles i.e., rotated or oriented in m ulti-dimensions, concerning how the image or video is captured. No matter the angle in which we capture the image of human action, whether it's a still or an act of drinking, jumping, etc.

- ✓ **Illumination** :
  - The image or video captured of a human being might be taken at dif-ferent illumination or brightness level, which makes it difficult to identify. Our image classification system must be able to grasp the dissimilarity in illumination. So, when we give any image of the same human's action with different brightness levels to our image classification system, it should be able to identify the same action.

HUMAN FEATURE RECOGNITION BASED ANALYSIS FOR DIGITAL TWIN USING CNN MODEL

- It is important to extract features of human behavior in digital twin environment and to select all the significant human characteristics that enhance efficiency of the model of machine learning or deep learning.
- For these neural networks, the disappearing and exploding gradient is a common issue altogether.
- The backpropagation algorithm is linked to this problem.

Role of ANN in human action recognition

- As inputs are interpreted only in forward direction, ANN is referred to as a Feed-Forward Neural network. As a gaggle of multiple neurons in each sheet, ANN is taken into account. As shown in Figure 8.9, ANN has three layers: the input layer, the secret layer, and the output layer.

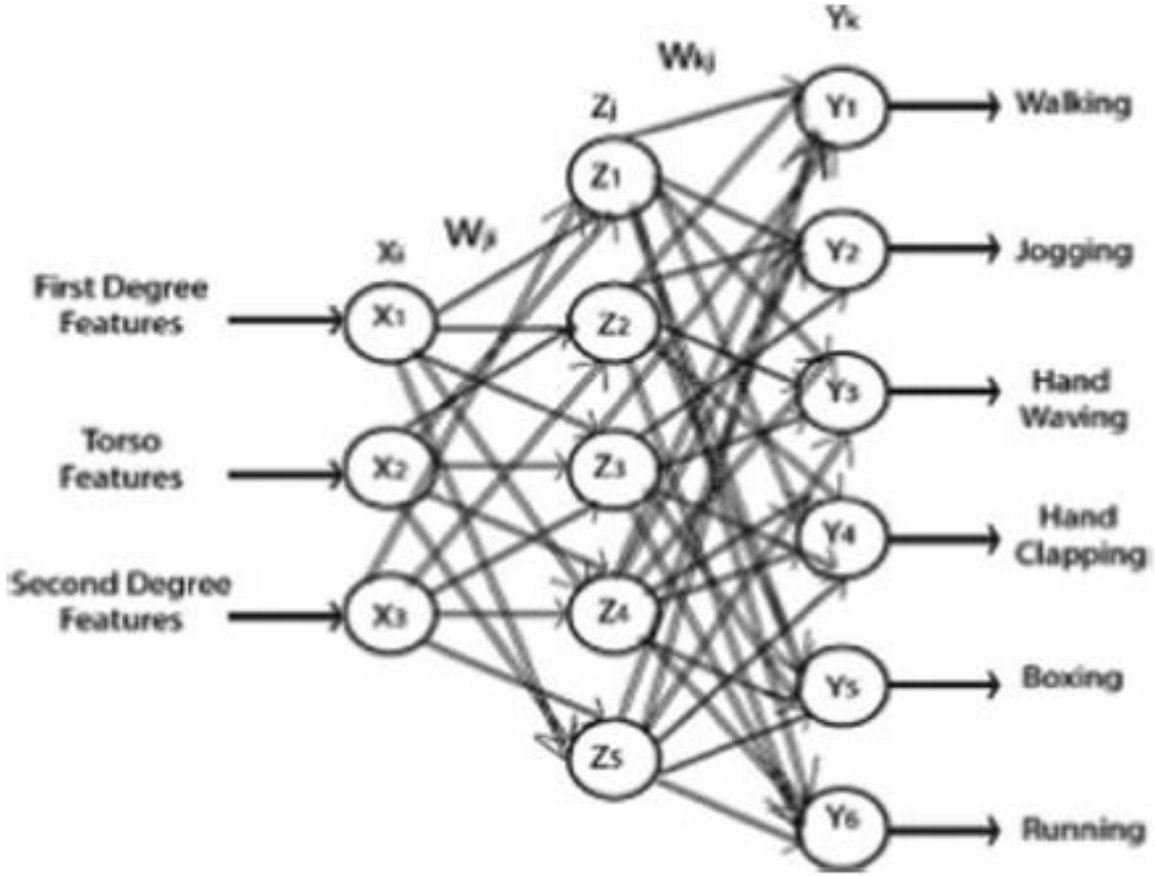


FIGURE 8.9 ANN structure [35].

Role of ANN in human action recognition

- The input layer accepts inputs, the hidden layer processes the inputs, and the result is then generated by the output layer. Each layer attempts to classify those weights.

- The Artificial Neural Network is good at learning any nonlinear function that enables the network to understand complex input-output relationship. The key move is to transform a two-dimensional image into a one-dimensional image, when solving a picture classification problem using ANN.

- The use of ANN has two drawbacks. First, the number of trainable parameters increase dramatically, with an increase in the image size, and second, ANN lacks a picture's spatial features that ask for the pixels in an image to be organized.

Role of CNN in human action recognition

- Convolutional neural networks (CNN) are all the craze within the deep learning community recently. These CNN models are getting used across different applications and domains. They are especially prevalent in image and video processing.

- The convolutional layer is composed of a set of convolutional kernels, where each neuron acts as a kernel. Although, if the kernel is symmetric, the convolution operation becomes a correlation operation. Convolutional kernel works by dividing the image into small slices, commonly known as receptive fields.

- The division of an image into small blocks helps in extracting feature motifs. Kernel convolves with the images using a specific set of weights by multiplying its elements with the corresponding elements of the receptive field.

- CNN can be built with several layers. Each neuron is connected with every other neuron of the next layer. Every layer gets its input from the output of the previous layer.

- There might be convolution layers, pooling layers, and fully connected layers, related to a common activation function. At last, there is often a SoftMax layer for classification. So, CNN is often imagined as a series of layers. Figure 8.10 shows the CNN architecture.

Role of CNN in human action recognition

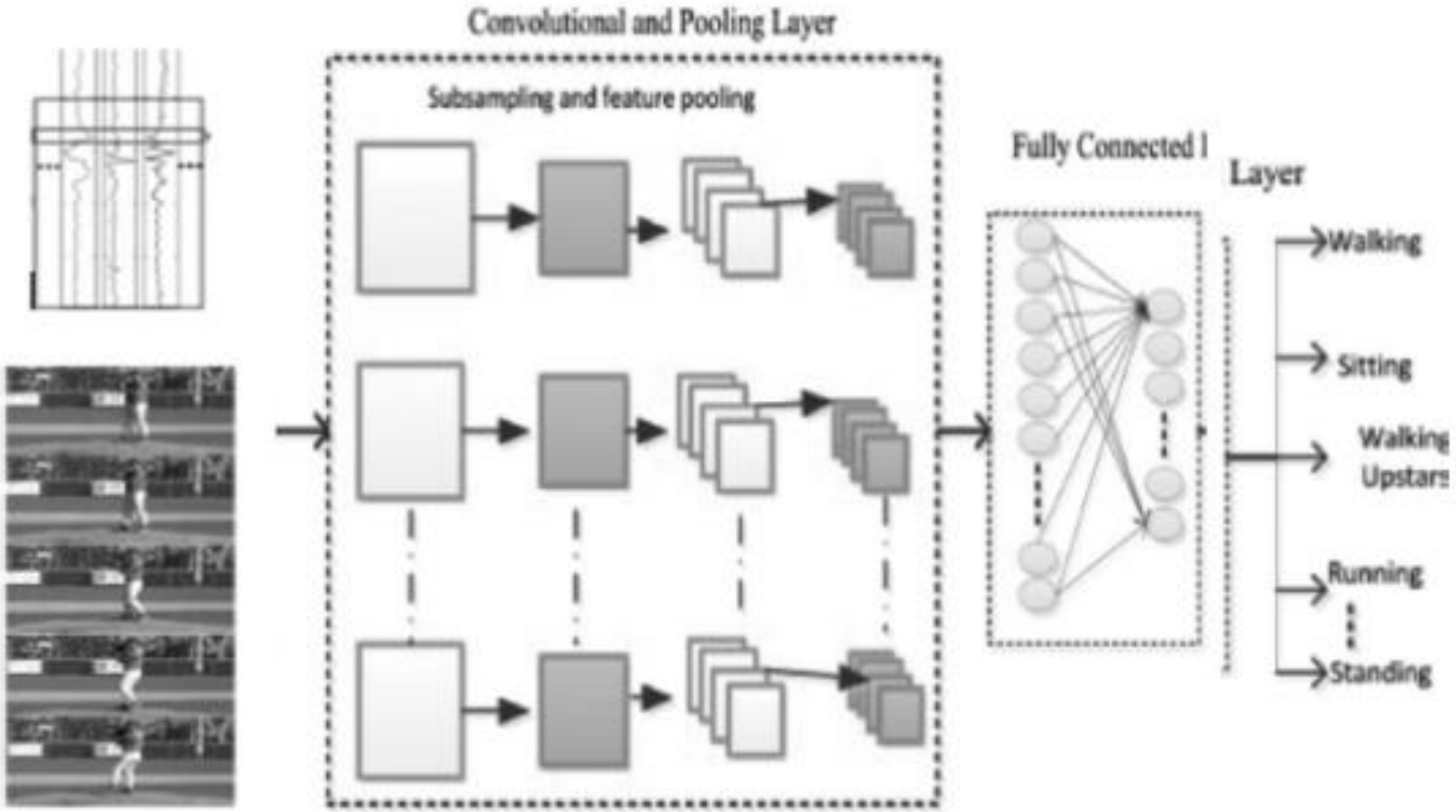


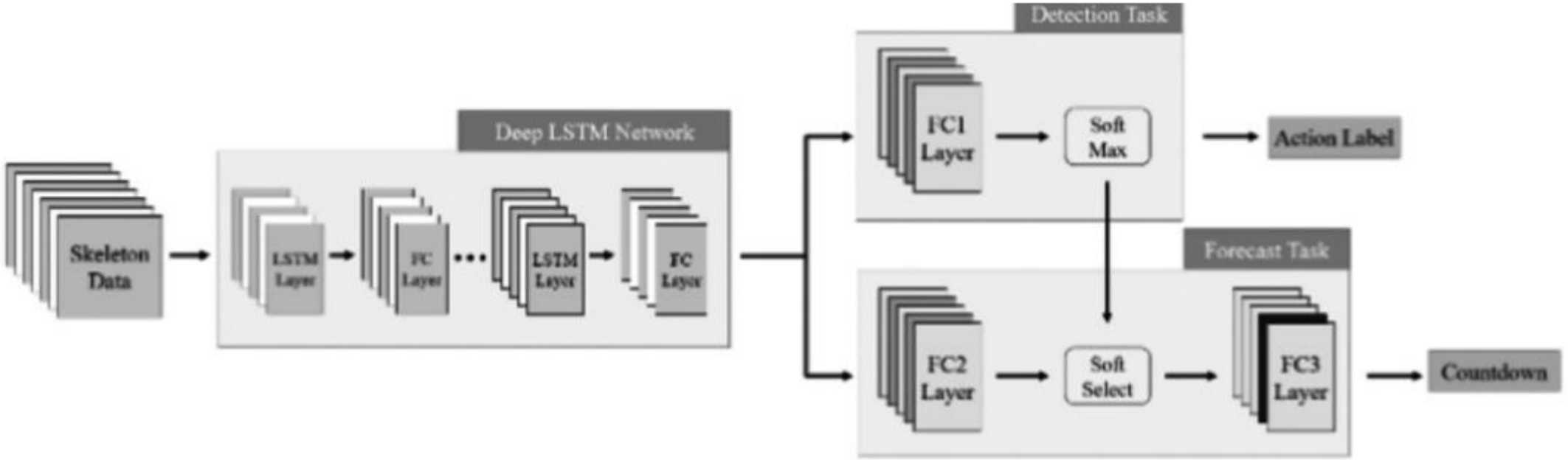
FIGURE 8.10 CNN Structure [41].

Role of CNN in human action recognition

- CNN learns to **filter automatically**, that is, without directly stating it. These filters assist in extracting from the input data of human behavior with the right and appropriate features.
- From an image, CNN captures the spatial features. Spatial characteristics are the groupings of pixels that accurately identify the link between the behavior of human beings and their surroundings.
- The location of a human being in a frame is often correctly identified by its association with another environment. The conception of parameter sharing is also followed by CNN.
- One filter is implemented as a function map through various parts of an input that needs to be supplied.

Role of RNN in human action recognition

- RNN captures dependency within the image while making predictions. It can also process both single data points, such as images, and sequences of data, such as speech, video, human activity, etc. RNNs share the parameters across different time pace shown in Figure 8.11.



Role of RNN in human action recognition

- It is often popularly referred to as Parameter Sharing. It leads to fewer parameters to train and reduce computational cost. D -RNNs are RNNs, with a huge time pace that also suffer from vanishing and exploding gradient problem.

- The RNN type of Long Short-Term Memory (LSTM) consists of a cell and three regulatory agencies in each unit, namely input and output gates, and a forget gate. The values of arbitrary time intervals are remembered by each cell, while the gates control the flow of information.

- These networks are suitable for time- series-based data classification, processing, and prediction. In learning, processing, and classifying such types of data, LSTMs have proven to excel. It’s not only a series of layers but also has stacked up time-pace layers.

Role of RNN in human action recognition

TABLE 8.3
Comparison between ANN, CNN, and RNN in Human Recognition [43]

	ANN	CNN	RNN
Data Type	Tabular data, sensor-based data	Mainly images but can be used for the sequential image too.	Works better on a series of image (video)
Performance	Less powerful	More powerful.	Less than CNN but better than ANN
Main advantages	Fault forbearance can work with insufficient data	Shares weight and have a high-performance rate in activity recognition from images	Predicts time orders and remembers every information
Disadvantages	Hardware dependent and unaccepted behavior of the network.	A large amount of training data is required and does not encode an individual's position and inclination.	A situation like a gradient vanishing and exploding gradient occurs.

APPLICATION AREA OF DIGITAL TWIN WITH HUMAN ACTION RECOGNITION

- **Monitoring Daily activities**

  - Earlier, human actions were evaluated by human operators, for example, security and surveillance processes or monitoring the health status of a patient. When the number of camera viewing devices and technological monitoring devices increased, so did the number.

  - The task of managing and controlling operators is demanding and expensive, as a ceaseless operation is now necessary. In the case of home treatment, certain activities will also not be financially viable for staff deployment. So, monitoring daily activities is now done automatically with the advancement of technology using deep learning and machine learning techniques.

  - These actions, when used in the digital twin environment, can help to manage the disaster, reduce consumption of resources, and many more.

APPLICATION AREA OF DIGITAL TWIN WITH HUMAN ACTION RECOGNITION

- **Digital twin in healthcare**

  - As a digital twin provides real-time analysis with equipment and other physical assets, they can revolutionize healthcare operations.

  - A digital twin of the patient organ with recognition of the doctor's action can allow the surgeon to practice procedure and be prepared, before the actual operation.

  - Also, the need for elderly care is growing rapidly as the Baby Boomer generation is getting retired.

  - A major goal of current research in human activity, monitoring and digital twin analysis, is to develop new technologies and applications for elderly care. Those applications could help prevent harm, such as, by detecting dangerous situations for older people.

APPLICATION AREA OF DIGITAL TWIN WITH HUMAN ACTION RECOGNITION

- **Impact in industries**

  - In the vision of Industry 4.0, tasks, and responsibilities among human employees and robots are shared.

  - Automated robots can't fully replace manual labor in the foreseeable future. It is difficult for robots to emulate the rational aptitude of humans. But, because of the advancements in sensor technology and data processing, IT-supported approaches for automated activity recognition and assessment are gaining significance.

  - Examples of human-machine interconnection are Multi-Touch Technology [48], Smartwatch [49], Deco Exoskeleton, and many more.

  - Now we have twins of each product that analyze the working and provide us a future prediction. For new products or factories, it is first tested in a virtual environment, to check the product's capability and changes are made accordingly before launching in the market. This reduces most of the risk and helps the industry to work efficiently and saves time and money

Large-scale government implementation to manage Disaster and anomalous activity

- By interpreting and understanding human activity, we can recognize and predict the occurrence of disaster and crime and help the government, police, or other agencies to react immediately.
- A digital twin predominant to identify human factors that impact successful implementation of port security measures. To reliably associate data for a particular person, a successful and efficient identification scheme must be established and maintained.
- This will allow the identification of disaster beforehand, according to the past decade's data provided to the machine.

Conclusion

- With the advancement of digital twin, human action recognition remains a significant issue in computer vision. Human action recognition has improved tremendously over the decades that helps in perceiving data from the surrounding into the digital twin layer.

- The problem still arises in identifying the action accurately, when confronting sensible landscapes, notwithstanding the characteristic intraclass variety and interclass similitude issue, which affects prediction of Digital twin.

- This paper presents some of the techniques and problems faced, which need to be improvised_further, so to improve the overall performance.

- As for the future direction, a more accurate algorithm for action recognition is also needed, that when it is applied on different data set, it shows the same accuracy level, with more efficient representation, and real-time operations still remaining open problems. This will allow the digital twin to make a better prediction to some extent.

Thank you

Digital Twin Techniques in Recognition of Human Action Using the Fusion of Convolutional Neural Network

PRESENTED BY: ABIR EL

2023-03-06