Vision-based Human Action Recognition using Machine Learning Techniques

by

Allah Bux

A thesis submitted in partial fulfillment for the
degree of Doctor of Philosophy

Supervisors: Prof. Plamen Angelov and Prof. Zulfiqar Habib

School of Computing and Communications

December 2017

Declaration of Authorship

I, Allah Bux, declare that the thesis titled, **"Vision-based Human Action Recognition using Machine Learning Techniques"** and the work presented in it are my own. I confirm that:

- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
- Detailed breakdown of the publications is presented in the first chapter of the thesis.

Signed:

Allah Bux

December 13, 2017

Abstract

The focus of this thesis is on automatic recognition of human actions in videos. Human action recognition is defined as automatic understating of what actions occur in a video performed by a human. This is a difficult problem due to the many challenges including, but not limited to, variations in human shape and motion, occlusion, cluttered background, moving cameras, illumination conditions, and viewpoint variations.

To start with, The most popular and prominent state-of-the-art techniques are reviewed, evaluated, compared, and presented. Based on the literature review, these techniques are categorized into handcrafted feature-based and deep learning-based approaches. The proposed action recognition framework is then based on these handcrafted and deep learning based techniques, which are then adopted throughout the thesis by embedding novel algorithms for action recognition, both in the handcrafted and deep learning domains.

First, a new method based on handcrafted approach is presented. This method addresses one of the major challenges known as "viewpoint variations" by presenting a novel feature descriptor for multiview human action recognition. This descriptor employs the region-based features extracted from the human silhouette. The proposed approach is quite simple and achieves state-of-the-art results without compromising the efficiency of the recognition process which shows its suitability for real-time applications.

Second, two innovative methods are presented based on deep learning approach, to go beyond the limitations of handcrafted approach. The first method is based on transfer learning using pre-trained deep learning model as a source architecture to solve the problem of human action recognition. It is experimentally confirmed that deep Convolutional

Neural Network model already trained on large-scale annotated dataset is transferable to action recognition task with limited training dataset. The comparative analysis also confirms its superior performance over handcrafted feature-based methods in terms of accuracy on same datasets.

The second method is based on unsupervised deep learning-based approach. This method employs Deep Belief Networks (DBNs) with restricted Boltzmann machines for action recognition in unconstrained videos. The proposed method automatically extracts suitable feature representation without any prior knowledge using unsupervised deep learning model. The effectiveness of the proposed method is confirmed with high recognition results on a challenging UCF sports dataset.

Finally, the thesis is concluded with important discussions and research directions in the area of human action recognition.

Acknowledgements

In the name of ALLAH, the Merciful, the Compassionate.

I would like to thank Almighty ALLAH for his countless blessings bestowed upon me during my PhD journey. Firstly, I would like to express my deep gratitude to my supervisor and mentor Prof. Plamen Angelov for supporting my research directions, which allowed me to explore new ideas in the field of computer vision and machine learning. I am indebted to him for his supervision, encouragement, motivation, and support throughout the years of my PhD. Secondly, I am very much grateful to my joint supervisor Prof. Zulfiqar Habib for providing continuous support, encouragement, and invaluable comments on my work. Indeed, his guidance helped me throughout the research work and in writing of the thesis.

I wish to express my gratitude to Prof. Zhiguo Ding and Dr. Manfred Lau for being the panelists of my PhD appraisal and for their insightful comments. I would like to thank my fellow doctoral students for their cooperation, feedback, discussions and friendship. Thanks also go to the administrative staff of Infolab21 for extending their support and assistance.

Nevertheless, I would like to gratefully acknowledge the support of my previous teachers and supervisors for supporting my choice of pursuing this PhD.

I would like to acknowledge the COMSATS Institute of Information Technology, Pakistan for providing me the scholarship for pursuing this PhD.

Last but not least, I would like to convey my deepest gratitude to my family for their sincere prayers, supports, and sacrifices during my PhD studies.

This thesis is dedicated to the loving memories of my Mom
with eternal love and appreciation.

Contents

List of Figures

List of Tables

Abbreviations

HAR	Human Activity Recognition
DBNs	Deep Belief Networks
AAL	Ambient Assisted Living
HCI	Human-Computer Interaction
HRI	Human-Robot Interaction
SIFT	Scale-Invariant Feature Transform
HOG	Histogram of Oriented Gradients
ESURF	Enhanced Speeded-Up Robust Features
LBP	Local Binary Pattern
CNNs	Convolutional Neural Networks
SVM-KNN	Support Vector Machines and K-Nearest Neighbor
UCF	University of Central Florida
ToF	Time-of-flight
SVM	Support Vector Machine
STIP	Space Time Interest Point
STISM	SpatialTemporal Implicit Shape Model
BOW	Bag-of-Words
STVs	Space-Time Volumes
MEI	Motion-Energy-Image
SURF	Speeded-Up Robust Features
HOF	Histogram of Optical Flow
MBH	Motion Boundary Histogram

TSRVFs	Transported-Square Root Vector Fields
PCA	Principal Component Analysis
GMM	Gaussian Mixture Model
SFV	Stacked Fisher Vector
FV	Fisher Vector
HMDB	Human Motion Database
SAX	Symbolic Aggregate Approximation
IXMAS	INRIA Xmas Motion Acquisition Sequences
NFS	Neuro-Fuzzy Systems
LLC	Locality-constrained Linear Coding
RBMs	Restricted Boltzmann Machines
DNNs	Deep Neural Networks
RNNs	Recurrent Neural Networks
ReLU	Rectifier Linear Unit
SFA	Slow Feature Analysis
FSTCN	Factorized Spatio-Temporal Convolutional Networks
SP-CNN	Stratified Pooling-based CNN
GP	Genetic Programming
DF	Deep Features
MF	Motion Features
SF	Static Features
HMM	Hidden Markov Model
DAG-SVM	Directed Acyclic Graph-Support Vector Machine
ECOC	Error-Correcting Output Codes
BTA	Binary Tree Architecture
RBF	Radial Basis Function
LOSO	Leave-One-Sequence-Out
FPS	Frames Per Second
CD	Contrastive Divergence

Chapter 1

Introduction

IN THIS CHAPTER

The topic of PhD thesis and its motivation are introduced in Section 1.1. Aims and objectives of the thesis are presented in Section 1.2, and brief overview is given in Section 1.3. The major contributions of the thesis are demonstrated in Section 1.5. Finally, the chapter is concluded with the list of publications related to the thesis in Section 1.6.

1.1 Introduction

Vision-based Human Activity Recognition (HAR) is an important research area in the field of computer vision and machine learning. The purpose of the HAR is to automatically understand what kind of action is performed in the video. This is really a difficult problem due to many challenges involved in HAR. These challenges include: occlusion, variation in human shape, and motion, cluttered backgrounds, stationary or moving cameras, different illumination conditions, and viewpoint variations. However, the intensity of these challenges may vary depending on the category of an activity under consideration. Generally, the activities fall into four categories, i.e., gestures, actions, interactions,

and group activities. This division is mainly based on the complexity and duration of the activities [3].

Gesture: A gesture is defined as a basic movement of the human body parts that carry some meaning. Head shaking, hand waving, and facial expressions are some good examples of gestures. Usually, a gesture takes a very short amount of time and its complexity is the lowest among the mentioned categories.

Action: An action is a type of an activity that is performed by a single person. In fact, it is a combination of multiple gestures (atomic actions). Some examples of actions are walking, running, jogging, and punching.

Interaction: It is a type of an activity performed by two actors. One actor must be a human and the other one may be a human or an object. Thus, it could be a human-human interaction or a human-object interaction. Fighting between two persons, hand shaking, and hugging are examples of a human-human interaction, while a person using an ATM, a person using a computer, and a person stealing a bag are examples of a human-object interaction.

Group Activity: This is the most complex type of activity. Certainly, it can be a combination of gestures, actions, and interactions. It involves more than two humans and a single or multiple objects. A group of people protesting, two teams playing a game, and a group meeting, are good examples of group activities.

As the title suggests, this thesis considers the **"gestures"** and **"actions"** for recognition. Therefore, the recognition of "interactions" and "group activities" are not covered under the scope of this thesis.

In recent years, HAR has drawn much attention of researchers around the globe due to its important applications in the real-world scenarios. This is a great motivating factor for the author of the thesis, to select it as a PhD topic. There are many applications of HAR, including, but not limited to the following:

Intelligent Video Surveillance: Traditional security surveillance systems use many cameras and require laborious human monitoring for video content analysis. On the other hand, intelligent video surveillance systems are aimed at automatically tracking the individuals or a crowd and recognizing their activities [4]. This includes the detection of suspicious or criminal activities and reporting to the authorities for immediate action [5]. In this way, the workload of the security personnel can be reduced and alerts can be signaled for security events, which can be helpful in preventing dangerous situations.

Ambient Assisted Living (AAL): The AAL is one of the important applications of HAR-based systems. These systems are used in health care to understand and analyze the patient's activities, to facilitate health workers in treatment, diagnosis, and general health care of the patients. In addition to this, these systems are also used for monitoring the daily life activities of elderly people, to provide them a safe, independent and comfortable stay. Usually, these systems capture the continuous movement of elderly people, automatically recognize their activities and detect any abnormality as it happens, such as falling down, having a stroke or respiration issues. Among these abnormal activities, "falls" are a major cause for fatal injury, especially for elderly people, and are considered a major hindrance for independent living [6].

Human-Computer Interaction (HCI): In addition to the conventional computer interfacing such as mouse and keyboard, it is desirable to have more natural interfaces between the computer and human operator by understanding the human gesture. An example of such a type of interfacing is controlling the presentation of slides using hand movement [7].

Human-Robot Interaction (HRI): This is also an important application of vision-based activity recognition. Giving a robot the ability to recognize human activities is very important for HRI. This makes the robots useful for the industrial setup and well as in the domestic environment as a personal assistant. In the domestic environment, one of the applications of HRI can be seen as humanoid robots that could recognize human emotions from the sequence of images [8]. In addition to this, the actor (robot) wearing the camera, may be involved in an ongoing activity. This is not only recognition of an activity in real time but also recognizing the activity before its completion [9].

Entertainment: Human activity recognition systems are used for recognition of entertainment activities such as dance [10], and sports [11]. The modelling of a player's action in the game has achieved much attention from the sports community in recent years due to its important use, such as adapting to the change in the game as it occurs [12].

Intelligent Driving: Human activity recognition techniques are also employed to assist drivers by providing different cues regarding the state of the driver while driving a vehicle. It has been reported that the secondary tasks performed by the drivers such as answering the phone, sending or receiving text messages, eating or drinking while operating a vehicle cause inattentiveness which may lead to accidents [13, 14].

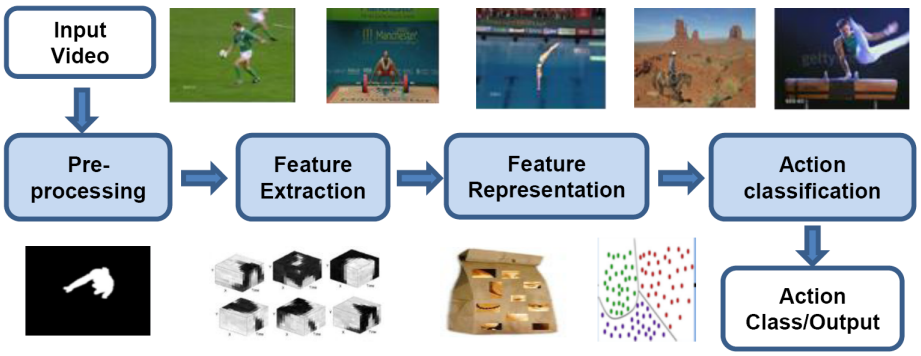


FIGURE 1.1: Schematic diagram of a typical activity recognition system

1.2 Aims and objectives of the thesis

During the last decade, different approaches have been proposed for HAR based on the design methodology and data collection process. These approaches fall into vision-based, non-visual sensor-based, and multi-modal categories. Indeed, HAR systems are complex in nature, and comprised of different sub-systems. The schematic diagram of a typical activity recognition system including its subsystems is presented in Figure 1.1. This thesis is aimed at developing novel techniques by adopting vision-based approach to human action recognition using handcrafted and deep learning-based techniques. Detailed objectives of the thesis are as follows:

- O1. comprehensive review of the state-of-the-art techniques based on handcrafted and deep learning approaches, in order to decide, which one works the best;
- O2. understanding the limitations of state-of-the-art techniques, and identify the gaps for contributions;

- O3. development of a novel method for view-invariant human action recognition, which is considered as one of the major challenges for recognition of human actions in different application domains;
- O4. development of an innovative method for human action recognition using a supervised deep learning or a transfer learning model;
- O5. development of an innovative method for human action recognition using an unsupervised deep learning model;
- O6. comparison between the supervised and unsupervised deep learning models on a same dataset;
- O7. production of better results than the existing ones in terms of accuracy and efficiency on standard benchmark datasets.

1.3 Overview of the thesis

This thesis consists of 6 chapters including the current introductory chapter of the thesis. Chapter 2 to chapter 4 are based on the already published work in international journals and conferences, while the work presented in chapter 5 is currently under review for journal publication. Each of these chapters start with a brief introduction explaining the context of the chapter and its contribution. Chapter 6 concludes the thesis and presents realization of each objective. Moreover, it also presents extension and enhancement of the proposed methods as future research directions. A brief overview of each chapter is presented below:

Chapter 2. Literature Review

This chapter presents a comprehensive review of the state-of-the-art techniques for HAR based on handcrafted and deep learning-based approaches, offering comparison, analysis, and discussions on these approaches. In addition to this, well-known public datasets available for experimentations are also presented to provide further insight into the field. Finally, the chapter is concluded with important discussions and research directions.

Chapter 3. Human Action Recognition from Multiple Views

This chapter presents a novel feature descriptor for multiview human action recognition. This descriptor employs the region-based features extracted from human silhouettes. To achieve this, a human silhouette is divided into regions in a radial fashion with the interval of a certain degree, and then region-based geometrical and Hu-moments features are obtained from each radial bin to articulate the feature descriptor. A multiclass support vector machine classifier is used for action classication. The proposed approach is quite simple and achieves state-of-the-art results without compromising the efciency of the recognition process. The contribution is two-fold. Firstly, this method achieves high recognition accuracy using a simple silhouette-based representation. Secondly, the average testing speed of the proposed method is 34 frames per second which is much higher than the existing methods and shows its suitability for the real-time applications. The extensive experiments on a well-known multiview IXMAS (INRIA Xmas Motion Acquisition Sequences) dataset confirmed the superior performance of the proposed method as compared to similar state-of-the-art methods.

Chapter 4. Human Action Recognition Using Transfer Learning

This chapter presents an innovative method for human action recognition using a pre-trained deep Convolutional Neural Networks (CNNs) model as a source architecture for extracting features from the target dataset. Then, a hybrid Support Vector Machines and K-Nearest Neighbour (SVM-KNN) classifier is used for action classification. It is experimentally confirmed that a pre-trained CNN based representations on a large-scale annotated dataset are transferable to an action recognition task with limited training dataset. The comparative analysis confirmed that the proposed method outperforms the handcrafted feature-based methods in terms of accuracy on the same datasets.

Chapter 5. Human Action Recognition Using Deep Belief Networks

This chapter presents an unsupervised Deep Belief Networks (DBNs) based method for recognizing human activities in unconstrained videos. This method automatically extracts suitable feature representation without any prior knowledge using an unsupervised deep learning model. In this work, the contribution is twofold. Firstly, the proposed method achieves high recognition accuracy as compared to state-of-the art deep learning and handcrafted feature-based methods. It demonstrates the potential of less explored unsupervised DBNs architecture for modelling complex human activities. Secondly, the proposed model uses and automatic structure learning method for learning important parameters. Moreover, although the model has been designed by keeping in mind the human activities, however, this model can also be adapted for other visual recognition tasks as well. The effectiveness of the proposed method is confirmed by high recognition results on a challenging UCF sports dataset.

Chapter 6. Conclusion and Future Research Directions

This chapter concludes the thesis, and presents the Comparative analysis and key findings regarding all proposed methods. In addition to this, possible extensions of the proposed methods both in short-term and long-term perspectives are presented in this chapter.

1.4 Validation methodology and software/hardware tools

The process of validation is aimed at assessing the correctness of the proposed algorithms. In this research work, the experimental methodology is used to evaluate and validate the performance of the proposed methods. In this regard, experiments have been conducted on publicly available benchmark datasets of human action recognition, including simple and complex datasets. The simple datasets such as KTH [15], in which the videos were recorded with static camera and background, and complex dataset such as UCF sports [16, 17], collected from variety of sports, broadcasted on television channels such as BBC and ESPN. In a UCF sports dataset, actions were recorded in real sport environment exhibiting the variations in background, illumination conditions, and occlusions, which make it a challenging dataset.

For experimentations, the MATLAB with its different versions (R2015, R2016, and R2017) was used as base software for implementation of the proposed algorithms. The hardware used for experimentation include Intel Core i7-4770 CPU with 8GB RAM, and Intel Xeon processor E5-2630, 64GB RAM with NVIDIA Quadro K2200 4GB GPU, for handcrafted and deep learning-based methods, respectively.

1.5 Contributions of the thesis

This section presents major contributions of the thesis.

1. **A comprehensive review and analysis of handcrafted and deep learning based approaches for HAR.**

   This thesis provides a comprehensive review, comparison, analysis, and evaluation of the state-of-the-art techniques based on handcrafted and deep learning approaches. By identifying limitations of state-of-the-art, this thesis proposed a framework which encompasses novel methods presented in the different chapters of the thesis. In addition to this, well-known public datasets and important applications of HAR are also presented to provide further insight into the field. This review covers all these aspects of HAR with comprehensive coverage of each part.

   The Human activity recognition contribution is presented in Chapter 2, and has also been published [18, 19].

2. **A method for human action recognition from multiple views based on view-invariant feature descriptor using support vector machines (SVMs).**

   This method is based on a novel view invariant feature descriptor, and can recognize the actions from five different views including, front, back, left, right and top view. This is achieved by extracting region-based features from a human silhouette. The proposed approach is efficient and achieves high recognition accuracy which makes it suitable for the real-time applications.

   This contribution is presented in Chapter 3, and has also been published [20].

3. **A method for applying a hybrid classifier and transfer learning to action recognition.**

   Supervised deep learning models have produced state-of-the-art results in object recognition, where a huge amount of labeled data are available. However, these models do not generalize well when the size of the dataset is small. In this direction, an innovative transfer learning-based method for HAR using hybrid classifier is proposed. This method uses a pre-trained Convolutional Neural Networks (CNNs) model as a source architecture for extracting features from the target dataset i.e., HAR dataset. It is experimentally confirmed that we can benefit from the pre-trained deep learning models rather than training the deep model from scratch in the case of a small dataset. Moreover, It is also confirmed that hybrid classifier yields superior performance than a single classifier with transfer learning.

   This contribution is presented in Chapter 4, and has also been published [21].

4. **A method for human action recognition using DBNs with automatic structure learning.**

   The importance of unsupervised deep learning models is obvious because most of the video content available on the internet is unlabeled. In this direction, an unsupervised Deep Belief Networks (DBNs) based method with automatic structure learning is proposed for HAR. The contribution of this method is two fold: Firstly, the proposed method achieved higher accuracy than the similar state-of-the-art methods, which confirms the potential of a less explored unsupervised DBNs architecture for modelling complex human activities. Secondly, the proposed method

uses an automatic structure learning method for learning important parameters. This helped in automatic extraction of suitable feature representation without any prior knowledge using an unsupervised model, where labels were introduced at fine-tuning stage only. Moreover, although the model was designed by keeping in mind the human activities, however, this model can also be adapted for other visual recognition tasks.

This contribution is presented in Chapter 5, currently, it is under review for journal publication.

1.6 Publications related to the thesis

In this section, the list of publications related to the thesis is presented.

Journal Papers

1. **Allah Bux Sargano,**, Plamen Angelov, and Zulfiqar Habib. Human Action Recognition from Multiple Views Based on View-Invariant Feature Descriptor Using Support Vector Machines.Applied Sciences 6.10 (2016): 309. **[Citation: 4]**

2. **Allah Bux Sargano**, Plamen Angelov, and Zulfiqar Habib. A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition. Applied Sciences 7.1 (2017): 110. **[Citation: 5]**

3. **Allah Bux Sargano**, Plamen Angelov, and Zulfiqar Habib. Human Action Recognition in Realistic Videos Using Deep Belief Networks, [**Under review**].

Conference Papers

1. **Allah Bux Sargano**, Plamen Angelov, and Zulfiqar Habib. Vision based human activity recognition: a review. Advances in Computational Intelligence Systems. Springer International Publishing, 2017. 341-371. **[Citation: 4]**

2. **Allah Bux Sargano**, Xiafeng Wang, Plamen Angelov, and Zulfiqar Habib. Human Action Recognition using Transfer Learning with Deep Representations. International Joint Conference on Neural Networks (IJCNN) 2017, USA . **Citation: 0**

Chapter 2

Literature Review

IN THIS CHAPTER

A comprehensive review of the handcrafted and learning based HAR techniques is presented with detailed introduction of these approaches presented in Section 2.1. While detailed comparisons, evaluations, and important discussions on handcrafted-based techniques are given in Section 2.2. Learning based approaches such as dictionary learning and deep learning based techniques are described, compared, and evaluated in Section 2.3. Important public datasets available for experimentations and evaluation of HAR techniques are presented in Section 2.4. Finally, the chapter is concluded in Section 2.5.

2.1 Introduction

Human Activity Recognition (HAR) plays an important role in many real-world applications. Its goal is to recognize the activities of a person or a group of persons from the sensors and/or video data, including the knowledge of the context in which these activities take place. Due to the advancement in sensor and visual technology, HAR based systems have been widely used in many real-world applications. Specifically, the proliferation of

small size sensors have enabled the smart devices to recognize the human activities in a context-aware manner [22]. Based on the design methodology and data collection process, these approaches are categorized into visual sensor-based, non-visual sensor-based, and multi-modal categories. The major difference between the visual and other types of sensors is the way of perceiving the data. Visual sensors provide the data in the form of 2D or 3D images or videos, whereas other sensors provide the data in the form of a one-dimensional signal [23].

In recent years, the wearable devices have been featured with many small non-visual sensors, which have enabled the development of pervasive applications. The wearable devices such as smart-phones, smart-watches, and fitness wristbands are worn all day long by many people. These devices have computing power, communication capability, and are available at low cost, which make them suitable for HAR [24]. Currently, various techniques have been proposed for sensor-based human activity recognition in daily health monitoring, rehabilitative training, and disease prevention [25].

On the other hand, visual sensor-based approach is one of the most popular HAR approach in the computer vision and machine learning research community. This approach has been employed in a wide range of application domains. In particular, the past decade has witnessed enormous growth in its applications. The major applications of vision-based HAR include: Human Computer Interaction (HCI), intelligent video surveillance, ambient assisted living, human-robot interaction, entertainment, and content-based video search. In HCI, the activity recognition systems observe the task carried out by the user

and guide him/her to complete it by providing feedback. In video surveillance, the activity recognition system can automatically detect a suspicious activity and report it to the authorities for immediate action. Similarly, in entertainment, these systems can recognize the activities of different players in the game.

Multi-modal HAR approach has also become popular during the last decade. In this approach, visual and non-visual sensors are used at the same time to recognize the human activities. This approach is specifically useful in the situations where one type of sensor is not enough to meet the user requirements. For example, a visual sensor, like camera can cover the subject and the context in which the activity take place but it may not be enough to analyze the sensitive information such as temperature, user heart rate, and humidity in the environment [25]. To overcome, these limitations, multi-modal approach is employed.

However, non-visual sensors in general and wearable sensors in specific have several limitations. Most of the wearable sensors need to be worn and run continuously, which may be difficulty to implement in real-world application scenarios due to many practical and technical issues. The major practical issues are acceptability and willingness to use wearable sensors and technical issues include battery life, ease of use, size, and effectiveness of the sensor [26]. In addition to this, in some application domains such as video surveillance where continuous monitoring of the people is required for suspicious activities, non-visual sensor-based approach might not be effective. Therefore, the ultimate solution lies in adopting vision-based human activity recognition approach as it can be applied to most of the application domains. This is the rationale for the proposed work

to focus on the vision-based human activity recognition. Depending on the complexity and duration, vision-based activities fall into four categories, i.e., gestures, actions, interactions, and group activities [3, 18] as shown in Figure 2.1.

Since the 1980s, researchers have been working on HAR from images and videos. One of the important directions that researchers have been following for action recognition is similar to the working of the human vision system. At a low level, the human vision system can receive the series of observations regarding the movement and shape of the human body in a short span of time. Then, these observations are passed to the intermediate human perception system for further recognition of the class of these observations, such as walking, jogging, and running. In fact, the human vision and perception system is robust and very accurate in recognition of observed movements. In order to achieve a similar level of performance by a computer-based recognition system, researchers invested much effort during the past few decades. However, unfortunately, due to many challenges and issues involved in HAR such as environmental complexities, intra-class variations, viewpoint variations, occlusions, and the non-rigid shape of the humans and objects, we are still far from the level of the human vision system. What we have achieved so far may be a fraction of what a mature human vision system can do.

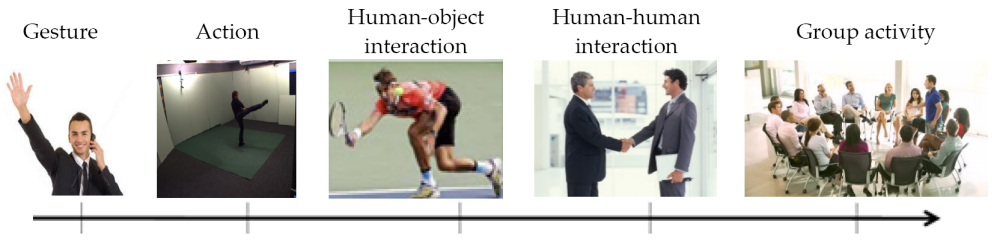


FIGURE 2.1: Categorization for different level of activities

Based on the comprehensive investigation of the literature, vision-based human activity recognition approaches can be divided into two major categories. (1) The traditional handcrafted representation-based approach, which is based on the expert designed feature detectors and descriptors such as Hessian3D, Scale-Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG), Enhanced Speeded-Up Robust Features (ESURF), and Local Binary Patterns (LBPs). This is followed by a generic trainable classifier for action recognition as shown in Figure 2.2. (2) The learning-based representation approach, which is a recently emerged approach with capability of learning features automatically from the raw data. This eliminates the need of handcrafted feature detectors and descriptors required for action representation. Unlike the traditional handcrafted approach, it uses the concept of a trainable feature extractor followed by a trainable classifier, introducing the concept of end-to-end learning, as shown in Figure 2.3.

The handcrafted representation-based approach mainly follows the bottom-up strategy for HAR. Generally, it consists of three major phases (foreground detection, handcrafted feature extraction and representation, and classification) as shown in Figure 2.4. A good number of survey papers have been published on different phases of handcrafted representation-based HAR processes in which different taxonomies have been used to discuss the HAR approaches. One survey presented in the work of Aggarwal and Ryoo

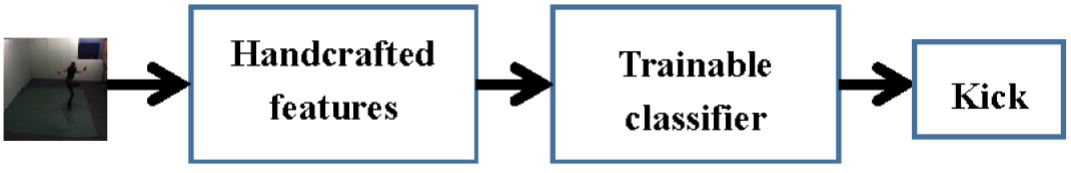


FIGURE 2.2: Example of a kicking action using handcrafted representation-based approach

[3], divides the activity recognition approaches into two major categories: single layered approaches and hierarchical approaches. Single-layered approaches recognize the simple activities from the sequence of a video, while hierarchical approaches recognize more complex activities by decomposing them into simple activities (sub-events). These are further sub-categorized such as space-time volumes, and trajectories, based on the feature representation and classification methods used for recognition. A detailed survey on object segmentation techniques is presented [27], discussing the challenges, resources, libraries and public datasets available for object segmentation. Another study, presented in [28], discussed the three levels of HAR, including core technology, HAR systems, and applications. Activity recognition systems are significantly affected by the challenges such as occlusion, anthropometry, execution rate, background clutter, and camera motion as discussed in [29]. This survey categorized the existing methods based on their abilities for handling these challenges on which potential research areas were also identified [29]. In [30], human action recognition methods based on the feature representation and classification were discussed. Similarly, Weinland et al. [31] surveyed the human activity recognition methods by categorizing them into segmentation, feature representation and classification. Ziaeefard and R. Bergevin produced a review article on semantic-based human action recognition methods [32], presenting the state-of-the-art methods for activity recognition which use semantic-based features. In this paper, semantic space, and semantic-based features such as pose, poselet, related objects, attributes, and scene context are also defined and discussed. Different handcrafted features extraction and representation methods have been proposed for human action recognition [33–37].

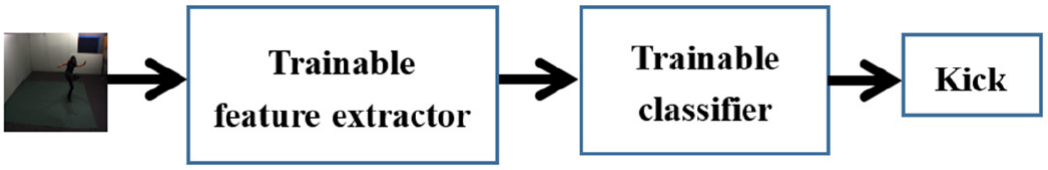


FIGURE 2.3: Example of a kicking action using learning-based representation approach

On the other hand, a learning-based representation approach, specifically, deep learning uses computational models with multiple processing layers based on representation learning with multiple levels of abstraction. This learning encompasses a set of methods that enable the machine to process the data in raw form and automatically transform it into a suitable representation needed for classification. This is what we call trainable feature extractors. This transformation process is handled at different layers, for example, an image consists of an array of pixels, and then the first layer transforms it into edges at a particular location and orientation. The second layer represents it as collection of motifs by recognizing the particular arrangement of edges in an image. The third layer may combine the motifs into parts and the following layers would turn it into the recognizable objects. These layers are learned from the raw data using a general purpose learning procedure which does not need to be designed manually by the experts [38]. This paper further examines various computer-based fields such as 3D games and animations systems [39, 40], physical sciences, health-related issues [41–43], natural sciences and industrial academic systems [44, 45].

One of the important components of vision-based activity recognition system is the camera/sensor used for capturing the activity. The use of appropriate cameras for capturing the activity has a great impact on the overall functionality of the recognition system. In fact, these cameras have been instrumental to the progression of research in the field of

computer vision [46–50]. According to the nature and dimensionality of images captured by these cameras, they are broadly divided into two categories, i.e., 2D and 3D cameras. The objects in the real world exist in 3D form: when these are captured using 2D cameras then one dimension is already lost, which causes the loss of some important information. To avoid the loss of information, researchers are motivated to use 3D cameras for capturing the activities. For the same reason, 3D-based approaches provide higher accuracy than 2D-based approaches but at higher computational cost. Recently, some efficient 3D cameras have been introduced for capturing images in 3D form. Among these, 3D Time-of-flight (ToF) cameras, and Microsoft Kinect have become very popular for 3D imaging. However, these sensors also have several limitations such as these sensors only capture the frontal surfaces of the human and other objects in the scene. In addition to this, these sensors also have a limited range of about 67 m, and data can be distorted by scattered light from the reflective surfaces [51].

Recently, thermal cameras have become popular. These are the passive sensors that capture the infrared variations emitted by the objects with temperature above absolute zero. Originally, these cameras were developed for surveillance as a night vision tool for military, but due to the significant reduction in prices, these cameras have become affordable for many applications. Deploying these cameras in computer vision systems can overcome the limitations of normal grayscale and RGB cameras such as illumination problems. In human activity recognition, these cameras can easily detect the human motion regardless of illumination conditions and colors of human surfaces and backgrounds. These cameras can also overcome another major challenge in human activity recognition

known as cluttered backgrounds [52, 53]. However, there is no any universal rule for selecting the appropriate camera; it mainly depends on the nature of the problem and its requirements.

A good number of survey and review papers have been published on HAR and related processes. However, due to the great amount of work published on this subject published reviews are quickly out-of-date. For the same reason, writing a review paper on human activity recognition is hard work and a challenging task. In this chapter we provide the discussion, comparison, and analysis of state-of-the-art methods of human activity recognition based on both handcrafted and learning-based action representations along with well-known datasets. This chapter covers all these aspects of HAR in a single chapter with reference to the more recent publications. However, the major focus remained on human gesture, and action recognition techniques as this is the motive of the thesis.

2.2 Handcrafted Representation-Based Approach

The traditional approach for activity recognition is based on the handcrafted feature-based representation. This approach has been popular among the HAR community and has achieved remarkable results on different public well-known datasets. In this approach, the important features from the sequence of image frames are extracted and the feature descriptor is built up using expert designed feature detectors and descriptors. Then, classification is performed by training a generic classifier such as Support Vector Machine (SVM) [54]. This approach includes space-time, appearance-based, local binary patterns, and fuzzy logic-based techniques as shown in Figure 2.4.

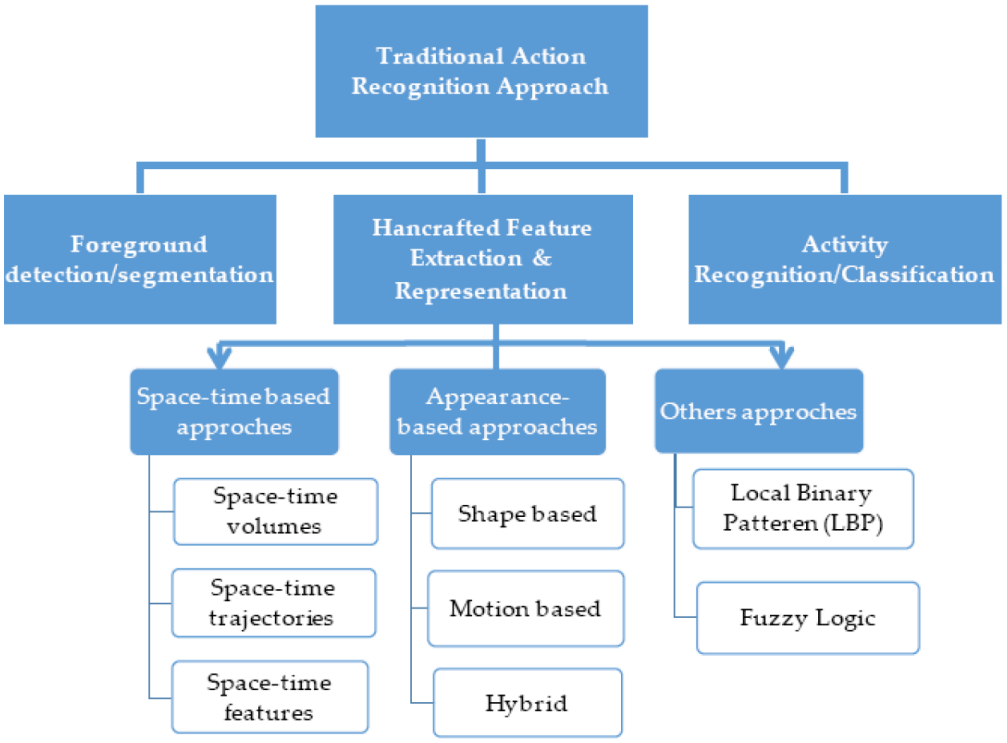


FIGURE 2.4: Traditional action representation and recognition approach

2.2.1 Space-Time-Based Approaches

Space-time-based approaches have four major components: space time interest point (STIP) detector, feature descriptor, vocabulary builder, and classifier [55]. The STIP detectors are further categorized into dense and sparse detectors. The dense detectors such as V-FAST, Hessian detector, dense sampling, densely cover all the video content for detection of interest points, while sparse detectors such as cuboid detector Harris3D [56], and Spatial Temporal Implicit Shape Model (STISM), use a sparse (local) subset of this content. Various STIP detectors have been developed by different researchers [57, 58]. The feature descriptors are also divided into local and global descriptors. The local descriptors such as cuboid descriptor, Enhanced Speeded-Up Robust Features (ESURF), and N-jet are based on the local information such as texture, colour, and posture, while

global descriptors use global information such as illumination changes, phase changes, and speed variation in a video. The vocabulary builders or aggregating methods are based on bag-of-words (BOW) or state-space model. Finally, for the classification, a supervised or unsupervised classifier is used, as shown in Figure 2.5.

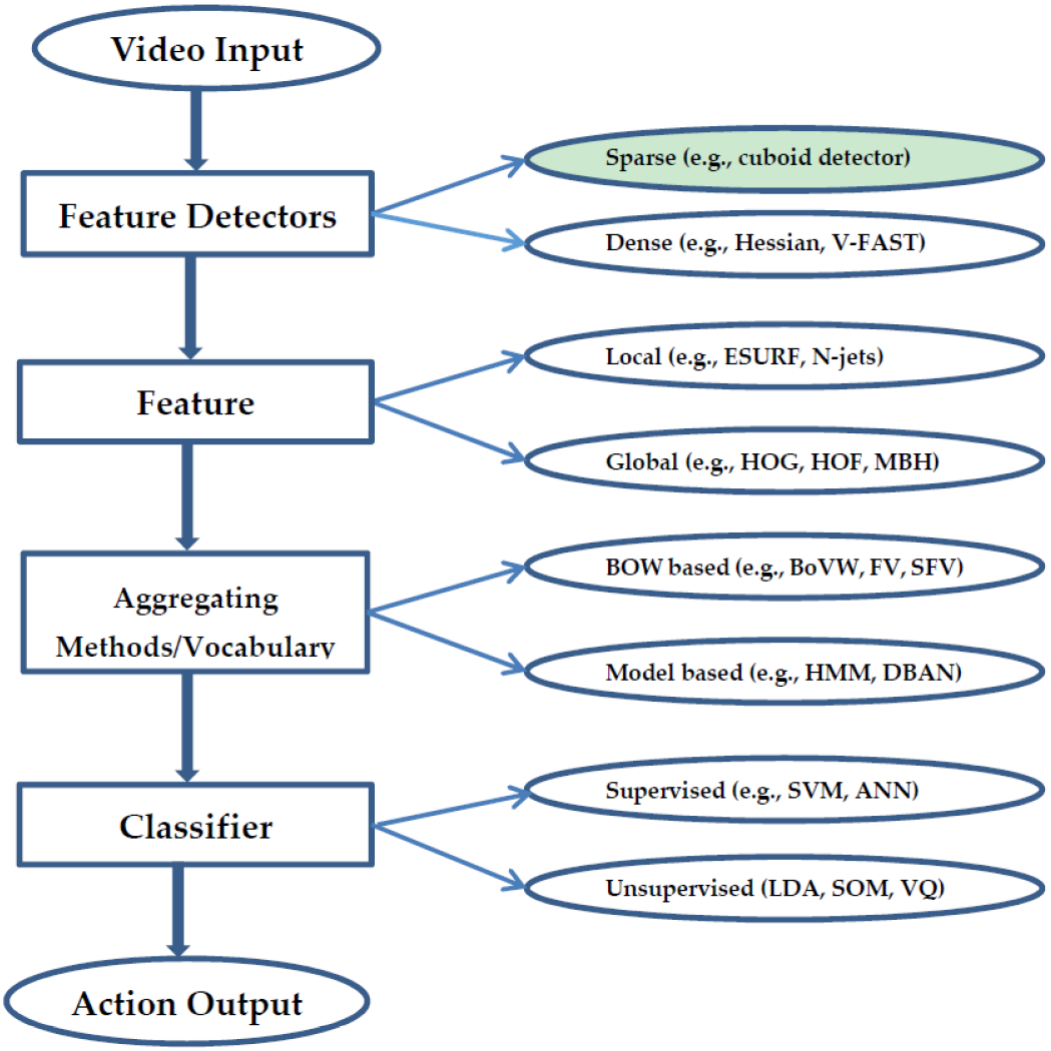


FIGURE 2.5: Components of space-time-based approaches

2.2.1.1 Space-Time Volumes (STVs)

The features in space-time domain are represented as 3D spatio-temporal cuboids, called space-time volumes (STVs). The core of STVs-based methods is a similarity measure between two volumes for action recognition. Bobick and Davis [59], proposed an action recognition system using template matching, instead of using space-time volumes, they used templates composed of 2D binary motion-energy-image (MEI) and motion-history-image (MHI) for action representation followed by a simple template matching technique for action recognition. Hu et al. [60]extended this work where MHI and two appearance-based features namely foreground image and histogram of oriented gradients (HOG) were combined for action representation, followed by simulated annealing multiple instance learning support vector machine (SMILE-SVM) for action classification. The method proposed by Roh et al. [61] also extended the work of Bobick and Davis[59] from 2D to 3D space for view-independent human action recognition using a volume motion template. The comparative analysis of these techniques is presented in Table 2.1.

2.2.1.2 Space-Time Trajectory

Trajectory-based approaches interpret an activity as a set of space-time trajectories. In these approaches, a person is represented by 2-dimensional (XY) or 3-dimensional (XYZ) points corresponding to his/her joints position of the body. When a person performs an action, there are certain changes in his/her joint positions according to the nature of the action. These changes are recorded as space-time trajectories, which construct a 3D XYZ or 4D XYZT representation of an action. The space-time trajectories work by tracking the

TABLE 2.1: Comparison of Space-Time-based approaches for HAR

Method	**Feature Type**	**Accuracy(%)**
	KTH [15]	
Sadanand and Corso 2012 [62]	Space-time volumes	98.2
Wu et al. 2011 [63]	Space-time volumes	94.5
Ikizler and Duygulu 2009 [64]	Space-time volumes	89.4
Peng et al. 2013 [65]	Features	95.6
Liu et al. 2011 [66]	Features(Attributes)	91.59
Chen et al. 2015 [67]	Features (mid-level)	97.41
Wang et al. 2011 [68]	Dense trajectory	95
	UCF Sports [16, 17]	
Sadanand and Corso 2012 [69]	Space-time volumes	95.0
Wu et al. 2011 [63]	Space-time volumes	91.30
Ma et al. 2015 [70]	Space-time volumes	89.4
Chen et al. 2015 [67]	Features (mid-level)	92.67
Wang et al. 2013 [71]	Features (Pose-based)	90
	HDMB-51 [72]	
Wang and Schmid 201 [73]	Dense trajectory	57.2
Jiang et al. 2012 [74]	Trajectory	40.7
Wang et al. 2011 [68]	Dense trajectory	46.6
Kliper et al. 2012 [75]	Space-time volumes, bag-of-visual-words	29.2
Sadanand and Corso 2012 [69]	Space-time volumes	26.9
Kuehne et al. 2011 [72]	Features	23.0
Wang et al. 2013 [76]	Features (mid-level)	33.7
Peng et al. 2014 [77]	Fisher vector and Stacked Fisher Vector	66.79
Jain et al. 2013 [78]	Features	52.1
Fernando et al. 2015 [79]	Features (Video Darwin)	63.7
Hoai and Zisserman 2014 [80]	Features	65.9
	Hollywood2 [81]	
Wang and Schmid 2013 [73]	Dense trajectory	64.3
Jain et al. 2013 [78]	Trajectory	62.5
Jiang et al. 2012 [74]	Trajectory	59.5
Vig et al. 2012 [82]	Trajectory	59.4
Mathe and Sminchisescu 2012 [83]	Space-time volumes	61.0
Kihl et al. 2016 [84]	Features	58.6
Lan et al. 2015 [85]	Features (mid-level)	66.3
Fernando et al. 2015 [79]	Features (Video Darwin)	73.7
Hoai and Zisserman [80]	Features	73.6
	Microsoft Research Action3D [86]	
Wang et al. 2013 [71]	Features (pose-based)	90.22
Amor et al. 2016 [87]	Trajectory	89
Zanfir et al. 2013 [88]	3D Pose	91.7
	YouTube action dataset [89]	
Wang et al. 2011 [68]	Dense trajectory	84.1
Peng et al. 2014 [77]	Features (FV + SFV)	93.38

joint position of the body in order to distinguish different types of actions. Following this idea many approaches have been proposed for action recognition based on the trajectories [71, 90, 91].

Inspired by the dense sampling in image classification, the concept of dense trajectories for action recognition from videos was introduced [68]. The authors sampled the dense points from each image frame and tracked them using displacement information from a dense optical flow field. These types of trajectories cover the motion information and are robust to irregular motion changes. This method achieved state-of-the-art-results with challenging datasets. In the work of Wang and Schmid [73], an extension to the work of Wang et al. [68] was proposed for the improvement of performance regarding camera motion. For estimation of camera motion the authors used Speeded-Up Robust Features (SURF) descriptor and dense optical flow. This significantly improved the performance of motion-based descriptors such as histograms of optical flow (HOF), and motion boundary histograms (MBH). However, when incorporating high density with trajectories within the video, it increases the computational cost. Many attempts have been made to reduce the computational cost of the dense trajectory-based methods. For this purpose, a saliency-map to extract the salient regions within the image frame was used in the work of Vig et al [82]. Based on the saliency-map a significant number of dense trajectories can be discarded without compromising the performance of the trajectory-based methods.

Recently, a human action recognition method from depth movies captured by the Kinect sensor was proposed [87]. This method represents dynamic skeleton shapes of the human body as trajectories on Kendalls shape manifold. This method is invariant to execution

rate of the activity and uses transported-square root vector fields (TSRVFs) of trajectories and standard Euclidean norm to achieve the computational efficiency. Another method used different descriptors such as HOG, HOF and motion boundary histograms (MBH) for recognition of actions of construction workers using dense trajectories [92]. Among these descriptors, authors reported the highest accuracy with codebook of size 500 using a MBH descriptor. Human action recognition in unconstrained videos is a challenging problem and few methods have been proposed at this end. For this purpose, a human action recognition method was proposed using explicit motion modelling [93]. This method used visual code words generated from the dense trajectories for action representation without using the foreground-background separation method.

2.2.1.3 Space-Time Features

The space-time features-based methods extract features from space-time volumes or space-time trajectories for human action recognition. Generally, these features are local in nature and contain discriminative characteristics of an action. According to the nature of space-time volumes and trajectories, these features can be divided into two categories: sparse and dense. The features detectors that are based on interest point detectors such as Harris3D [57], and Dollar [94] are considered as sparse, while feature detectors based on optical flow are considered as dense. These interest point detectors provide the base for most of the recently proposed algorithms. In the work of Thi et al [95], the interest points were detected using Harris3D [57], based on these points they build the feature descriptor and used PCA (principal component analysis)-SVM for classification. In the work of Kihl et al [84], the authors proposed a novel local polynomial space-time descriptor based on

optical flow for action representation.

The most popular action representation methods in this category are based on the Bag-of-Visual-Words (BoVW) model [96, 97] or its variants [98, 99]. The BoVW model consists of four steps, feature extraction, code-book generation, encoding/pooling, and normalization. They extracted local features from the video; learn visual dictionary by using a training set by Gaussian Mixture Model (GMM) or K-mean clustering, encode and pool features, and finally represent the video as normalized pooled vectors followed by a generic classifier for action recognition. The high performance of the BoVW model is due to an effective low level feature such as dense trajectory features [73, 100], encoding methods such as Fisher Vector [99], and space-time co-occurrence descriptors [65]. The improved dense trajectory (iDT) [73] provides the best performance among the space-time features on several public datasets.

The coding methods have played an important role in boosting the performance of these approaches. Recently, a new encoding method named Stacked Fisher Vector (SFV) [77] was developed as an extension of traditional single layer Fisher Vector (FV) [99]. Unlike traditional FV, which encodes all local descriptors at once, SFV first performs encoding in dense sub-volumes, then compresses these sub-volumes into FV, and finally applies another FV encoding based on the compressed sub-volumes. For the detailed comparison of a single layer FV and stacked FV readers are encouraged to refer to [77].

2.2.2 Discussion

Space-Time-based approaches have been evaluated by many researchers on different well-known datasets including simple and complex activities as recorded in Table 2.1.

Some merits of these approaches are as follows: (1) STVs-based approaches are suitable for recognition of gestures and simple actions. However, these approaches have also produced comparable results on complex datasets such as Human Motion database (HMDB-51), Hollywood2, and University of Central Florida (UCF-101); (2) The space-time trajectory-based approaches are especially useful for recognition of complex activities. With the introduction of dense trajectories, these approaches have become popular due to their high accuracy for challenging datasets. In recent years, trajectory-based approaches have been getting lot of attention due to their reliability under noise and illumination changes; (3) Space-time feature-based approaches have achieved state-of-the art results on many challenging datasets. It has been observed that descriptors such as HOG3D, HOG, HOF, and MBH are more suitable for handling intra-class variations and motion challenges in complex datasets as compared to local descriptors such as N-jet.

However, these approaches have some limitations as follows: (1) STVs-based approaches are not effective in recognizing multiple persons in a scene; these methods use a sliding window for this purpose which is not very effective and efficient; (2) Trajectory-based approaches are good at analyzing the movement of a person in a view invariant manner but to correctly localize the 3D XYZ joint position of a person is still a challenging task; (3) Space-time features are more suitable for simple datasets; for effective results on complex datasets, a combination of different features is required which raises the computational complexity. These limitations can cause a hindrance to real-time applications.

2.2.3 Appearance-Based Approach

In this section, 2D (XY) and 3D (XYZ) depth image-based methods which use effective shape, motion, or combination of shape and motion features are discussed. The 2D shape-based approaches [101, 102] use shape and contour-based features for action representation and motion-based approaches Efros et al and Fathi and Mori [103, 104] use optical flow or its variants for action representations. Some approaches use both shape and motion feature for action representation and recognition [105]. In 3D-based approaches, a model of a human body is constructed for action representation; this model can be based on cylinders, ellipsoids, visual hulls generated from silhouettes or surface mesh. Some examples of these methods are 3D optical flow [106], shape histogram [107], motion history volume [108], and 3D body skeleton [109].

2.2.3.1 Shape-Based Approach

The shape-based methods capture the local shape features from the human image or silhouette [110]. These methods first obtained the foreground silhouette from an image frame using foreground segmentation techniques. Then, they extracted the features from the silhouette itself (positive space) or from the surrounding regions of the silhouette (negative space) between canvas and the human body [111]. Some of the important features that can be extracted from the silhouette are contour points, region-based features, and geometric features. The region-based human action recognition method was proposed in [112]. This method divides the human silhouette into a fixed number of grids and cells for action representation and used a hybrid classifier Support Vector Machine and Nearest Neighbour (SVM-NN) for action recognition. For practical applications,

the human action recognition method should be computationally lean. In this direction, an action recognition method was proposed using Symbolic Aggregate approximation (SAX) shapes [113]. In this method, a silhouette was transformed into time-series and these time-series were converted into a SAX vector for action representation followed by a random forest algorithm for action recognition. A pose-based view invariant human action recognition method was proposed based on the contour points with a sequence of the multiview key poses for action representation [114]. An extension of this method was proposed by Chaaraoui and Flrez-Revuelton[115]. This method uses the contour points of the human silhouette and radial scheme for action representation and support vector machine as a classifier. A region-based descriptor for human action representation was developed by extracting features from the surrounding regions (negative space) of the human silhouette [111, 116]. Another method used pose information for action recognition [117]. In this method, first, the scale invariant features were extracted from the silhouette, and then these features were clustered to build the key poses. Finally, the classification was performed using a weighted voting scheme.

2.2.3.2 Motion-Based Approach

Motion-based action recognition methods use motion features for action representation followed by a generic classifier for action recognition. A novel motion descriptor was proposed [118] for multiview action representation. This motion descriptor is based on motion direction and histogram of motion intensity followed by the support vector machine for classification. Another method based on 2D motion templates using motion history images and histogram of oriented gradients was proposed by Murtaza et al [119].

An action recognition method was proposed by Kliper-Gross et al [75], based on the key elements of motion encoding and local changes in motion direction encoded with the bag-of-words technique.

2.2.3.3 Hybrid Approach

These methods combine shape-based and motion-based features for action representation. An optical flow and silhouette-based shape features were used for view invariant action recognition in [120] followed by principal component analysis (PCA) for reducing the dimensionality of the data. Some other methods base on shape and motion information were proposed for action recognition [121, 122]. The coarse silhouette features, radial grid-based features and motion features were used for multiview action recognition in the work of Pelikan et al [122]. Meanwhile, Jiang et al [105] used shape-motion prototype trees for human action recognition. The authors represented action as a sequence of prototypes in shape-motion space and used distance measure for sequence matching. This method was tested on five public datasets and achieved state-of-the-art results. Eweiwi [123], proposed a method based on action key poses as a variant of Motion Energy Images (MEI), and Motion History Images (MHI) for action representation followed by a simple nearest-neighbour classifier for action recognition.

2.2.4 Other Approaches

In this section, the two important approaches that do not fit under the headings of the above mentioned categories are discussed. These approaches include Local Binary Pattern (LBP), and fuzzy logic-based methods.

2.2.4.1 Local Binary Pattern (LBP)-Based Approach

Local binary pattern (LBP) [124] is a type of visual descriptor for texture classification. Since its inception, several modified versions of this descripto [125–127] have been proposed for different classification-related tasks in computer vision. A human action recognition method was proposed [128] based on LBP combined with appearance invariance and patch matching method. This method was tested on different public datasets and proved to be efficient for action recognition. Another method for activity recognition was proposed using LBP-TOP descriptor [129]. In this method, the action volume was partitioned into sub-volumes and feature histogram was generated by concatenating the histograms of sub-volumes. Using this representation, they encoded the motion at three different levels: pixel level (single bin in the histogram), region-level (sub-volume histogram), and global-level (concatenation of sub-volume-histograms). The LBP-based methods have also been employed for multiview human action recognition. A multiview human action recognition method was proposed [130] based on contour based pose features and uniform rotation-invariant LBP followed by SVM for classification. Recently, another motion descriptor named Motion Binary Pattern (MBP) was introduced for multiview action recognition [131]. This descriptor is a combination of Volume Local Binary Pattern (VLBP) and optical flow. This method was evaluated on multiview INRIA Xmas Motion Acquisition Sequences (IXMAS) dataset and achieved 80.55% recognition results.

2.2.4.2 Fuzzy Logic-Based Approach

Traditional vision-based human action recognition methods employ the spatial or temporal features followed by a generic classifier for action representation and classification. However, it is difficult to scale up these methods for handling uncertainty and complexity involved in real world applications. For handling these difficulties, fuzzy-based methods are considered as being better choice.A fuzzy-based framework was proposed for human action recognition based on the fuzzy log-polar histograms and temporal self-similarities for action representation followed by SVM for action classification [132]. The evaluation of the proposed method on two public datasets confirmed the high accuracy and its suitability for real world applications. Another method based on fuzzy logic was proposed [133], which utilized the silhouette slices and movement speed features as input to the fuzzy system, and employed the fuzzy c-means clustering technique to acquire the membership function for the proposed system. The results confirmed the better accuracy of the proposed fuzzy system as compared to non-fuzzy systems for the same public dataset.

Most of the human action recognition methods are view dependent and can recognize the action from a fixed view. However, a real-time human action recognition method must be able to recognize the action from any viewpoint. To achieve this, many state-of-the-art methods use data captured by multiple cameras. However, this is not a practical solution because calibration of multiple cameras in real world scenarios is quite difficult. The use of a single camera should be the ultimate solution for view invariant action recognition. Along these lines, a fuzzy logic-based method was proposed [134] for view invariant action recognition using a single camera. This method extracted human contours from

the fuzzy qualitative Poisson human model for view estimation followed by clustering algorithms for view classification as shown in Figure 2.6 The results indicate that the proposed method is quite efficient for viewing independent action recognition. Some methods based on neuro-fuzzy systems (NFS) have also been proposed for human gesture and action recognition [135, 136]. In addition to this, evolving systems [49, 137] are also very successful in behaviour recognition.

2.2.5 Discussion

In this section, a comparison is made between the appearance of LBP and fuzzy logic-based methods.These approaches are simple and have produced state-of-the-art results

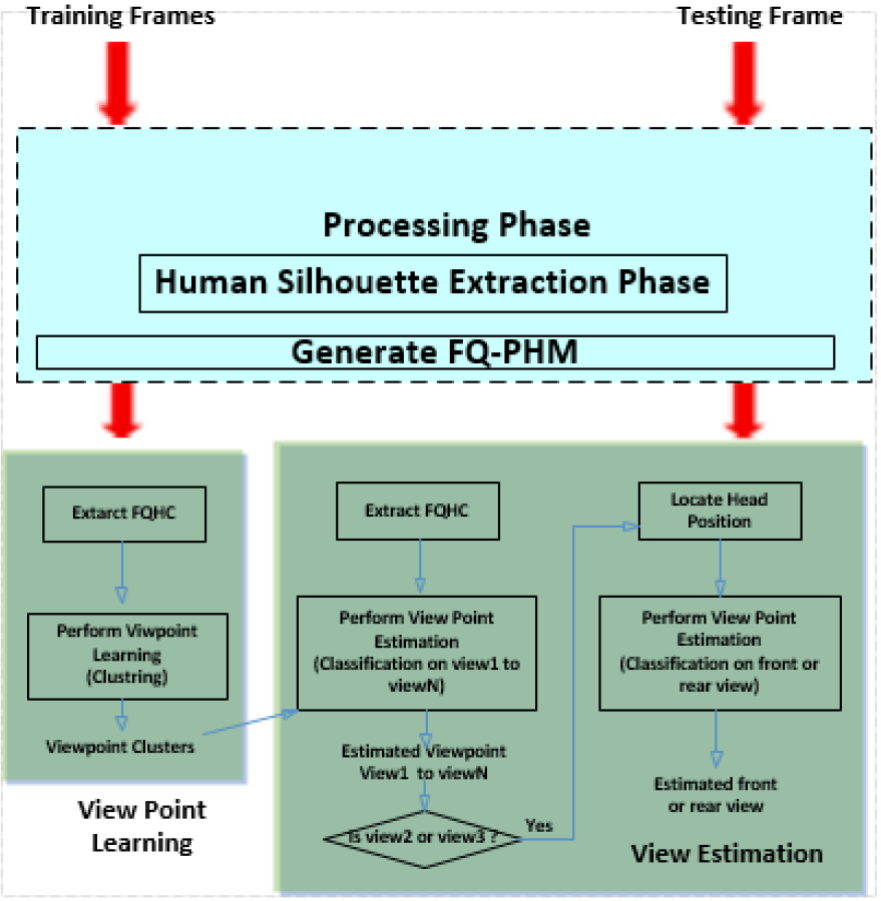


FIGURE 2.6: Example of Fuzzy view estimation framework

on Weizmann, KTH, and multiview IXMAS datasets as recorded in Table 2.2. There are two major approaches for multiview human action recognition based on shape and motion features: 3D approach and 2D approach [51]. As indicated in Table 2.2, 3D approaches provide higher accuracy than 2D approaches but at higher computational cost which makes these approaches less applicable to real time applications. In addition to this, it is difficult to reconstruct a good quality 3D model because it depends on the quality of extracted features or silhouettes of different views. Hence, the model is exposed to deficiencies which might have occurred due to segmentation errors in each view point. Moreover, a good 3D model of different views can only be constructed when the views overlap. Therefore, a sufficient number of viewpoints have to be available to reconstruct a 3D model.

2.3 Learning-Based Action Representation Approach

The performance of the human action recognition methods mainly depends on the appropriate and efficient representation of data. Unlike handcrafted representation-based approach where the action is represented by handcrafted feature detectors and descriptors; learning-based representation approaches have the capability to learn the feature automatically from the raw data, thus introducing the concept of end-to-end learning which means transformation from pixel level to action classification. Some of these approaches are based on a evolutionary approach (genetic programming) and dictionary learning while others employ deep learning-based models for action representation. We have divided these approaches into two categories: non-deep learning-based approach

TABLE 2.2: Comparison of appearance, Local Binary Pattern, and Fuzzy logic-based approaches

Method	Feature Type	Performance (%)
Weizmann [10]		
Rahman et al. 2012 [111]	Shape Features	100
Vishwakarma and Kapoor 2015 [112]	Shape Features	100
Rahman et al. 2014 [116]	Shape-motion	95.56
Chaaraoui et al. 2013 [114]	Shape Features	92.8
Vishwakarma et al. 2016 [121]	Shape Features	100
Jiang et al. 2012 [105]	Shape-motion	100
Eweiwi et al. 2011 [123]	Shape-motion	100
Yeffet and Wolf 2009 [128]	LBP	100
Kellokumpu et al. 2008 [129]	LBP (LBP-TOP)	98.7
Kellokumpu et al. 2011 [138]	LBP	100
Sadek et al. 2011 [132]	Fuzzy features	97.8
Yao et al. 2015 [133]	Fuzzy features	94.03
KTH [15]		
Rahman et al. 2012 [111]	Shape Features	94.67
Vishwakarma and Kapoor 2015 [112]	Shape Features	96.4
Rahman et al. 2014 [116]	Shape-motion	94.49
Vishwakarma et al. 2016 [121]	Shape Features	95.5
Sadek et al. 2012 [139]	Shape Features	93.30
Jiang et al. 2012 [105]	Shape-motion	95.77
Yeffet and Wolf 2009 [128]	LBP	90.1
Mattivi and Shao 2009 [140]	LBP (LBP-TOP)	91.25
Kellokumpu et al. 2011 [138]	LBP	93.8
Sadek et al. 2011 [132]	Fuzzy Features	93.6
IXMAS (INRIA Xmas Motion Acquisition Sequences) [141]		
Junejo et al. 2014 [113]	Shape features	89.0
Sargano et al. 2016 [20]	Shape features	89.75
Lin et al. 2009 [142]	Shape-motion	88.89
Chaaraoui et al. 2013 [114]	Shape features	85.9
Chun and Lee 2016 [118]	Motion features	83.03
Vishwakarma et al. 2016 [121]	Shape features	85.80
Holte et al. 2012 [143]	Motion feature (3D)	100
Weinland et al. 2006 [108]	Motion features (3D)	93.33
Turaga et al. 2008 [144]	Shape-motion (3D)	98.78
Pehlivan and Duygulu 2011 [145]	Shape features (3D)	90.91
Baumann et al. 2016 [131]	LBP	80.55

and deep learning-based approach as shown in Figure 2.7.

2.3.1 Non-Deep Learning-Based Approaches

These approaches are based on genetic programming and dictionary learning as discussed in the following section.

2.3.1.1 Dictionary Learning-Based Approach

Dictionary learning is a type of representation learning which is generally based on the sparse representation of the input data. The sparse representation is suitable for the categorization tasks in images and videos. Dictionary learning-based approaches have been employed in a wide range of computer vision applications such as image classification and action recognition [146]. The concept of dictionary learning is similar to Bag-of-visual-words(BoVW) model because both are based on the representative vectors learned from the large number of samples. These representative vectors are called code words, forming a code-book in BoVW model, and dictionary atoms in the context of dictionary

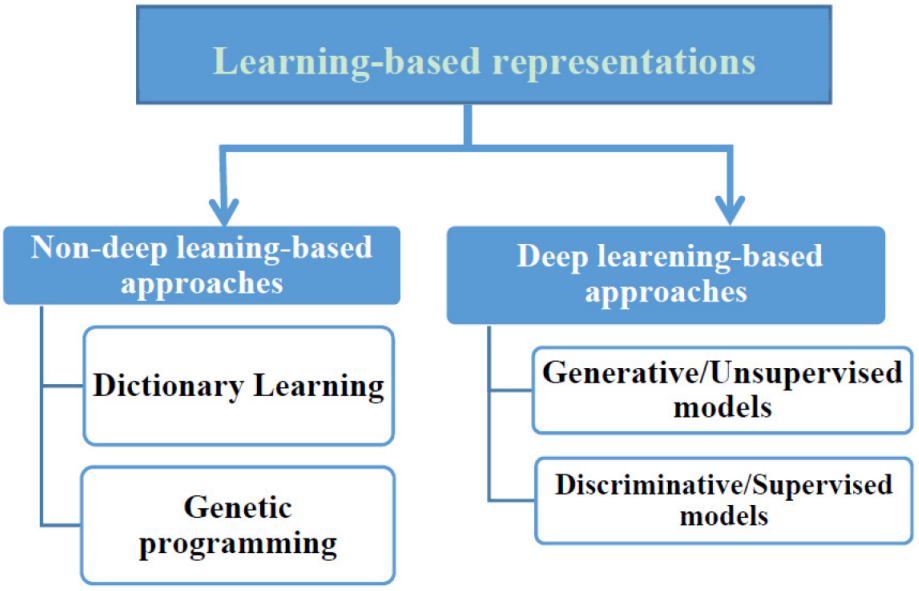


FIGURE 2.7: Learning-based action representation approaches

learning. One way to get the sparse representation of input data is to learn over-complete basis dictionary. In a paper published by Guha and Ward [147], three over-complete dictionary learning frameworks were investigated for human action recognition. An over-complete dictionary was constructed from a set of spatio-temporal descriptors, where each descriptor was represented by a linear combination of small number of dictionary elements for compact representation. A supervised dictionary learning-based method was proposed for human action recognition [148] based on the hierarchical descriptor. The cross-view action recognition problem was addressed by using transferable dictionary pair [149]. In this method the authors learned the view specific dictionaries where each dictionary corresponds to one camera view and extended this work by constructing a common dictionary which shares information from different views [150]. The proposed approach outperforms the state-of-the-art methods on similar datasets. A weakly supervised cross-domain dictionary learning-based method was proposed for visual recognition [151]. This method learns discriminative, domain-adaptive, and reconstructive dictionary pair and corresponding classifier parameters without any prior information.

Dictionary leaning-based methods also use unsupervised learning, for example, Zhu and Shao [152] proposed an unsupervised approach for cross-view human action recognition. This method does not require target view label information or correspondence annotations for action recognition. The set of low-level trajectory features are coded using locality-constrained linear coding (LLC) [153] to form the coding descriptors, then peak values are pooled to form a histogram that captures the local structure of each action.

2.3.1.2 Genetic Programming

Genetic programming is a powerful evolutionary technique inspired by the process of natural evolution. It can be used to solve the problems without having the prior knowledge of the solution. In human activity recognition, genetic programming can be employed to identify the sequence of unknown primitive operations that can maximize the performance of the recognition task. Recently, a genetic programming-based approach was introduced for action recognition [154]. In this method, instead of using handcrafted features, authors automatically learned the spatio-temporal motion features for action recognition. This motion feature descriptor evolved on population of 3D operators such as 3D Gabor filter and wavelet. In this way, the effective set of features was learned for action recognition. This method was evaluated on three challenging datasets and outperformed the handcrafted as well as other learning-based representations.

2.3.1.3 Bayesian Networks

A Bayesian network or belief network is a probabilistic graphical model, which represents the random variables and their dependencies in the form of directed acyclic graph. Different variations of Bayesian network have been introduced such as conditional Bayesian networks, temporal Bayesian networks, and Multi-Entity Bayesian Networks (MEBN). In the work of Zhnag et al [155], an Interval Temporal Bayesian Networks (ITBN) was introduced for recognition of complex human activities. This method combined the Bayesian network with the interval algebra to model the time dependencies over time intervals. In order to evaluate the performance of the proposed method, a cargo loading dataset

was considered for experimentations and evaluations. Khan et al [156] proposed another method for action detection using dynamic conditional Bayesian network, which also achieved the state-of-the-art results. In Park et al [157], MEBN was used for predictive situation awareness (PSAW) using multiple sensors. These networks are robust for reasoning the uncertainty in the complex domains for predicting and estimating the temporally evolving situations.

2.3.2 Deep Learning-Based Approach

Recent studies show that there are no universally best hand-crafted feature descriptors for all datasets, therefore learning features directly from the raw data may be more advantageous. Deep learning is an important area of machine learning which is aimed at learning multiple levels of representation and abstraction that can make sense of data such as speech, images, and text. Deep learning-based methods have ability to process the images/videos in their raw forms and automate the process of feature extraction, representation, and classification. These methods use trainable feature extractors and computational models with multiple processing layers for action representation and recognition. Based on a research study on deep learning presented [158], we have classified the deep learning models into three categories: (1) generative/unsupervised models (e.g., Deep Belief Networks (DBNs), Deep Boltzmann machines (DBMs), Restricted Boltzmann Machines (RBMs), and regularized auto-encoders); (2) Discriminative /Supervised models (e.g., Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs); (3) Hybrid models, these models use the characteristics of both models, for example a goal of discrimination may be assisted with the outcome of

the generative model. However, these models are not discussed here separately.

2.3.2.1 Generative/Unsupervised Models

Unsupervised deep learning models do not require the target class labels during the learning process. These models are specifically useful when labelled data are relatively scarce or unavailable. Deep learning models have been investigated since the 1960s [159] but researchers paid little attention these models. This was mainly due to the success of shallow models such as SVMs [160], and unavailability of huge amount of data, required for training the deep models.

A remarkable surge in the history of deep models was triggered by the work of Hinton et al [161] where the highly efficient DBN and training algorithm was introduced followed by the feature reduction technique [162]. The DBN was trained layer by layer using RBMs [163], the parameters learned during this unsupervised pre-training phase were fine-tuned in a supervised manner using back-propagation. Since the introduction of this efficient model there has been a lot of interest in applying deep learning models to different applications such as speech recognition, image classification, object recognition, and human action recognition.

A method using unsupervised feature learning from video data was proposed in the work of Le et al [164] for action recognition. The authors used an independent subspace analysis algorithm to learn spatio-temporal features combining them with deep learning techniques such as convolutional and staking for action representation and recognition. Deep Belief Networks (DBNs) trained with RBMs were used for human action recognition [165]. The proposed method outperformed the handcrafted learning-based approach

on two public datasets. Learning continuously from the streaming video without any labels is an important but challenging task. This issue was addressed in a paper by Hasan and Roy-Chowdhury [166] using an unsupervised deep learning model. Most of the action datasets have been recorded under a controlled environment; action recognition from unconstrained videos is a challenging task. A method for human action recognition from unconstrained video sequences was proposed by Ballan et al [167] using DBNs.

Unsupervised learning played a pivotal role in reviving the interests of the researchers in deep learning. However, it has been overshadowed by the purely supervised learning since the major breakthrough in deep learning used CNNs for object recognition [168].Conversely, an important study by the pioneers of latest deep learning models suggest that unsupervised learning is going to be far more important than its supervised counterpart in the long run [38] since we discover the world by observing it rather being told the name of every object. The human and animal learning is mostly unsupervised.

2.3.2.2 Discriminative/Supervised Models

According to the literature survey of human action recognition, the most frequently used model under the supervised category is Convolutional Neural Network (CNN). The CNN [169] is a type of deep learning model which has shown excellent performance at tasks such as pattern recognition, hand-written digit classification, image classification and human action recognition[168, 170]. This is a hierarchical learning model with multiple hidden layers to transform the input volume into output categories. Its architecture consists of three main types of layers: convolutional layer, pooling layer, and fully-connected layer as shown in Figure 2.8. Understanding the operation of the different

layers of CNN require mapping back these activities into pixel space, this is done with the help of Deconvolutional Networks (Deconvnets) [171]. The Deconvnets use the same process as CNN but in reverse order for mapping from feature space to pixel space. Initially, the deep CNN [169] was used for representation and recognition of objects from still images [168]. This was extended to action recognition from videos [172] using stacked video frames as input to the network but the results were worse than even the handcrafted shallow representations [73, 97]. This issue was investigated by Simonyan and Zisserman [2] who came up with the idea of two-stream (spatial and temporal) CNN for action recognition. An example of a two-stream convolutional neural network is shown in Figure 2.9. Both these streams were implemented as Convnet, the spatial stream recognizes the action from still video frames and the temporal stream performs action recognition from the motion in the form of dense optical flow. Afterwards, these two streams were combined using late fusion for action recognition. This method achieved superior results to one of the best shallow handcrafted-based representation methods [73]. However, the

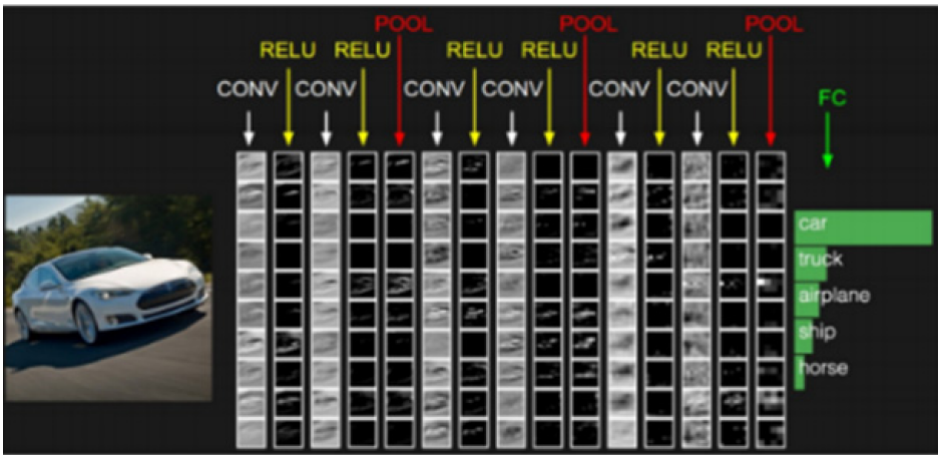


FIGURE 2.8: Different layers of Convolutional Neural Networks (Source [1])

two-stream architecture may not be suitable for real-time applications due to its computational complexity.

Most of the deep CNN models for action recognition are limited to handling inputs in 2D form. However, some applications do have data in 3D form that requires a 3D CNN model. This problem was addressed by Xu et al [173] by introducing the 3D convolutional neural networks model for airport surveillance. This model uses features from both spatial and temporal dimensions by performing 3D convolutions in the convolutional layer. This method achieved state-of-the-art results in airport video surveillance datasets. The supervised learning CNN model 2D or 3D can also be accompanied by some unsupervised endeavours. One of the unsupervised endeavours is slow feature analysis (SFA) [174], which extracts slowly varying features from the input signal in an unsupervised manner. Beside other recognition problems, it has also proved to be effective for human action recognition [175]. Sun et al [176], combined two-layered SFA learning with 3D CNN for automated action representation and recognition. This method achieved state-of-the-art results on three public datasets including KTH, UCF sports, and Hollywood2. Other types of supervised models include Recurrent Neural Networks (RNNs). A method

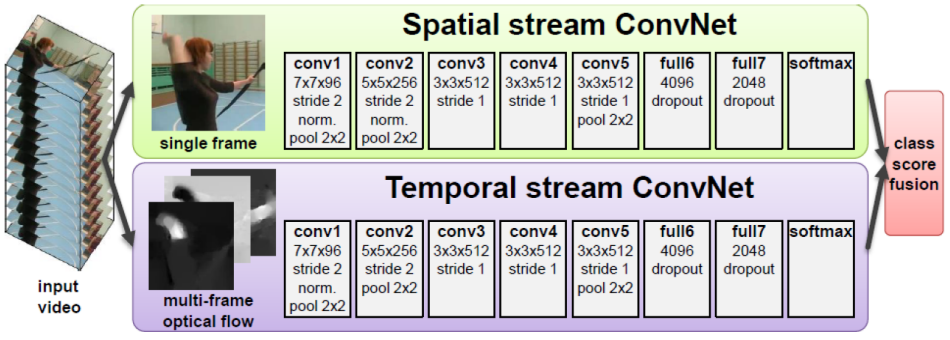


FIGURE 2.9: An example of a two-stream Convolutional Neural Network (CNN) architecture, source[2]

using RNN was proposed [177] for skeleton-based action recognition. The human skeleton was divided into five parts and then separately fed into five subnets. The results of these subnets were fused into the higher layers and final representation was fed into the single layer. For further detail regarding this model reader may refer to Du et al [177].

The deep learning-based models for human action recognition require a huge amount of video data for training. While collecting and annotating huge amount of video data is immensely laborious and requires huge computational resources. A remarkable success has been achieved in the domains of image classification, object recognition, speech recognition, and human action recognition using the standard 2D and 3D CNN models. However, there still exist some issues such as high computational complexity of training CNN kernels, and huge data requirements for training. To curtail these issues researchers have been working to come up with variations of these models. With this in mind, factorized spatio-temporal convolutional networks (FSTCN) were proposed by Sun et al [178] for human action recognition. This network factorizes the standard 3D CNN model as a 2D spatial kernel at lower layers (spatial convolutional layers) based on sequential learning process and 1D temporal kernels in the upper layers (temporal convolutional layers). This reduced the number of parameters to be learned by the network and thus reduced the computational complexity of the training CNN kernels and proposed its detailed architecture. Another approach using spatio-temporal features with a 3D convolutional network was proposed by Tran et al [179] for human action recognition. The evaluation of this method on four public datasets confirmed three important findings: (1) 3D CNN is more suitable for spatio-temporal features than 2D CNN; (2) The CNN architecture with small

3x3x3 kernels is the best choice for spatio-temporal features; (3) The proposed method with linear classifier outperforms the state-of-the-art methods.

Some studies have reported that incorporating handcrafted features into the CNN model can improve the performance of action recognition. Along this direction, combining information from multiple sources using CNN was proposed by Park et al [180]. In this method, handcrafted features were used to perform spatially varying soft-gating and used fusion method for combining multiple CNN trained on different sources. Recently, Yu et al [181] came up with the variation of CNN called stratified pooling-based CNN (SP-CNN). Since each video has a different number of frame-level features, to combine and get a video-level feature is a challenging task. The SP-CNN method addressed this issue by proposing variation in the CNN model as follows: (a) adjustment of pre-trained CNN on target dataset; (b) extraction of features at frame-level; (c) using principal component analysis (PCA) for dimensionality reduction; (d) stratified pooling frame-level features into video-level features; (e) SVM for multiclass classification. This architecture is shown in Figure 2.10.

Semantic-based features such as pose, poselet are important cues for describing the category of an action being performed. In this direction, some methods based on fuzzy CNN were proposed by Ijjina and Mohan and Choran et al [182, 183] using local posed-based features. These descriptors are based on the motion and appearance information acquired from tracking human body parts. These methods were evaluated on Human Motion Database (HMDB) produced superior results than other state-of-the-art methods. It

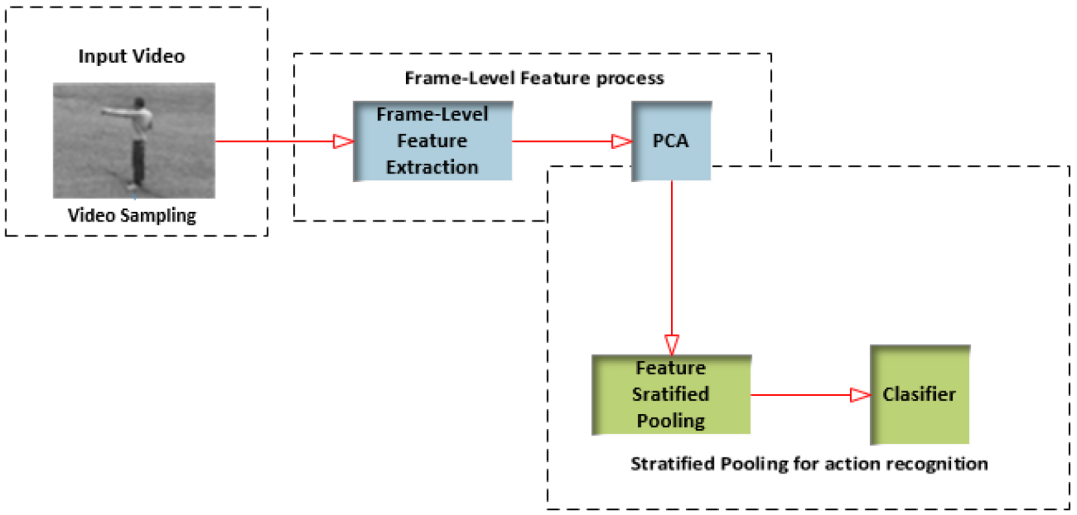


FIGURE 2.10: An example of stratified pooling with CNN

has been observed that the context/scene where the action is carried out also provides important cues regarding the category of an action. Gkioxari et al [184] used the contextual information for human action recognition and Girshick et al [185] adapted region-based Convolutional Neural Network (CNN) for classification. Where actor was considered as a primary region and contextual cues as a secondary region.

One of the major challenges in human action recognition is view variance. The same action viewed from different angles looks quite different. This issue was addressed by Rahmani and Mian [186] using CNN. This method generates the training data by fitting synthetic 3D human model to real motion and renders human poses from different viewpoints. The CNN model has shown better performance than handcrafted representation-based methods for multi-view human action recognition. The comparison of non-deep learning and deep learning-based methods on different public datasets is shown in Table 2.3.

TABLE 2.3: Comparison of Learning-based action representation approaches

Method	Feature Type	Performance (%)
	KTH [15]	
Wang et al. 2012 [148]	Dictionary Learning	94.17
Liu et al. 2016 [154]	Genetic Programming	95.0
Le et al. 2011 [164]	Subspace Analysis	93.9
Ballan et al. 2012 [167]	Codebook	92.66
Hasan and Chowdhury 2014 [166]	DBN (Deep Belief Networks)	96.6
Ji et al 2013 [173]	3D CNN (Convolutional Neural Networks	90.2
Zhang and Tao 2012 [175]	Slow Feature Analysis (SFA)	93.50
Sun et al. 2014 [176]	Deeply-Learned Slow Feature Analysis (D-SFA)	93.1
Alfaro et al. 2016 [187]	Sparse Coding	97.5
	HDMB-51 [72]	
Liu et al. 2016 [154]	Genetic Programming	48.4
Simonyan and Zisserman 2014 [2]	CNN	59.4
Luo et al. 2015 [188]	Actionness	56.38
Wang et al. 2015 [189]	Convolutional Descriptor	65.9
Lan et al. 2015 [190]	Multi-skip Feature Stacking	65.1
Sun et al. 2015 [178]	Spatio-Temporal CNN	59.1
Park et al. 2016 [180]	Deep CNN	54.9
Yu et al. 2016 [181]	SP (Stratified Pooling)-CNN	74.7
Bilen et al. 2016 [191]	Multiple Dynamic Images (MDI)	65.2
Mahasseni and Todorovic 2016 [192]	Long Short Term Memory (LSTM)-CNN	55.3
Fernando et al. 2016 [193]	Rank pooling + CNN	65.8
Zhu et al. 2016 [194]	Key Volume Mining	63.3
	Hollywood2 [81]	
Liu et al. 2016 [154]	Genetic Programming	46.8
Le et al. 2011 [164]	Subspace Analysis	53.3
Ballan et al. 2012 [167]	Codebook	45.0
Sun et al. 2014 [176]	DL-SFA	48.1
Fernando et al. 2016 [193]	Rank pooling + CNN	75.2
	MSR Action3D [86]	
Du et al. 2015 [177]	RNN (Recurrent Neural Network)	94.49
Wang et al. 2016 [195]	3D Key-Pose-Motifs	99.36
Veeriah et al. 2015 [196]	Differential RNN	92.03
	University of Central Florida (UCF-101) [188]	
Simonyan and Zisserman 2014 [2]	Two-stream CNN	88.0
Ng et al. 2015 [197]	CNN	88.6
Wang et al. 2015 [189]	Convolutional Descriptor	91.5
Lan et al. 2015 [190]	Multi-skip Feature Stacking	89.1
Sun et al. 2015 [178]	Spatio-Temporal CNN	88.1
Tran et al. 2015 [179]	3D CNN	90.4
Park et al. 2016 [180]	Deep CNN	89.1
Yu et al. 2016 [181]	SP-CNN	91.6
Bilen et al. 2016 [191]	MDI and Trajectory	89.1
Mahasseni and Todorovic 2016 [192]	LSTM-CNN	86.9
Zhu et al. 2016 [194]	Key volume mining	93.1
	UCF Sports [16, 17]	
Sun et al. 2014 [176]	DL-SFA	86.6
Weinzaepfel et al. 2015 [198]	Spatio-temporal	91.9%
	ActivityNet Dataset [199]	
Heilbron et al. 2015 [199]	Deep Features, and Motion Features (Untrimmed)	42.2
Heilbron et al. 2015 [199]	Deep Features, and Motion Features (Trimmed)	50.2

2.3.3 Discussion

In this section, learning-based action representation approaches are summarized. These approaches have been divided into genetic programming, dictionary learning, supervised and unsupervised deep learning-based approaches according to the learning representation used in each category. However, this division boundary is not strict and approaches may overlap.

The dictionary learning-based methods have attracted increasing interest of researchers in computer vision, specifically, in human activity recognition. These methods introduced the concept of unified learning of dictionary and corresponding classifier into a single learning procedure, which leads to the concept of end-to-end learning. On the other hand, genetic programming (GP) is a powerful evolutionary method inspired by natural selection, used to solve the problem without prior domain knowledge. In human action recognition, GP is used to design the holistic descriptors that are adaptive, and robust for action recognition. These methods have achieved state of the art results on challenging action recognition datasets.

Deep learning has emerged as highly popular direction within the machine learning which has outperformed the traditional approaches in many applications of computer vision. The highly advantageous property of deep learning algorithms is their ability to learn features from the raw data, which eliminates the need of handcrafted feature detectors and descriptors. There are two categories of deep learning models, i.e., unsupervised and supervised models. The DBN is a popular unsupervised model which has been used for human action recognition. This model has already achieved high performance on

challenging datasets as compared to its traditional handcrafted counterparts [176]. On the other hand, CNN is one of the most popular deep learning models in the supervised category. Most of the existing learning-based representations either directly apply CNN to video frames or variations of CNN for spatio-temporal features extraction and representation. These models have also achieved excellent results on challenging human activity recognition datasets as recorded in Table 2.3. So far, supervised deep learning models have achieved better performance but some studies suggest that unsupervised learning is going to be far more important in the long run. Since the world is discovered by observing it rather being told the name of every object, human and animal learning is mostly unsupervised [38].

The deep learning models have also some limitations: These models require huge amount of data for training the algorithms. Most of the action recognition datasets such as KTH [15], IXMAS [141], HMDB-51 [72], and UCF Sports [16, 17] are comparatively small for training these models. However, recently a large-scale ActivityNet dataset [199] was proposed with 200 action categories, 849 hours of video in total. This dataset is suitable to train deep learning-based algorithms. We can expect a major breakthrough with development of algorithms that could produce remarkable results on this dataset.

2.4 Datasets

In this section, well-known public datasets for human activity recognition are discussed. The focus is on recently developed datasets which have been frequently used for experimentations.

2.4.1 Weizmann Human Action Dataset

This dataset [10] was introduced by the Weizmann institute of Science in 2005. This dataset consists of 10 simple actions with static background, i.e., walk, run, skip, jack, jump forward or jump, jump in place or pjump, gallop-sideways or side, bend, wave1, and wave2. It is considered as a good benchmark for evaluation of algorithms proposed for recognition of simple actions. Some methods, such as [111, 112] have reported 100% accuracy on this dataset. The background of the dataset is simple and only one person performs the action in each frame as shown in Figure 2.11

2.4.2 KTH Human Action Dataset

The KTH dataset [15] was created by the Royal Institute of Technology, Sweden in 2004. This dataset consists of six types of human actions (walking, jogging, running,

FIGURE 2.11: One frame example of each action in Weizmann dataset

boxing, hand clapping and hand waving) performed by 25 actors with 4 different scenarios. Thus, it contains $25 \times 6 \times 4 = 600$ video sequences. These videos were recorded with static camera and background; therefore, this dataset is also considered relatively simple for evaluation of human activity recognition algorithms. The method proposed by Sadanand and Corso [69] achieved 98.2% accuracy on this dataset, which is the highest accuracy reported so far. The one frame example of each action from four different scenarios is shown in Figure 2.12.

2.4.3 IXMAS Dataset

INRIA Xmas Motion Acquisition Sequences (IXMAS) [141] a multiview dataset was developed for evaluation of view-invariant human action recognition algorithms in 2006. This dataset consists of 13 daily life actions performed by 11 actors 3 times each. These actions include crossing arms, stretching head, sitting down, checking watch, getting up, walking, turning around, punching, kicking, waving, picking, pointing, and throwing. These actions were recorded with five calibrated cameras including 4 side cameras and

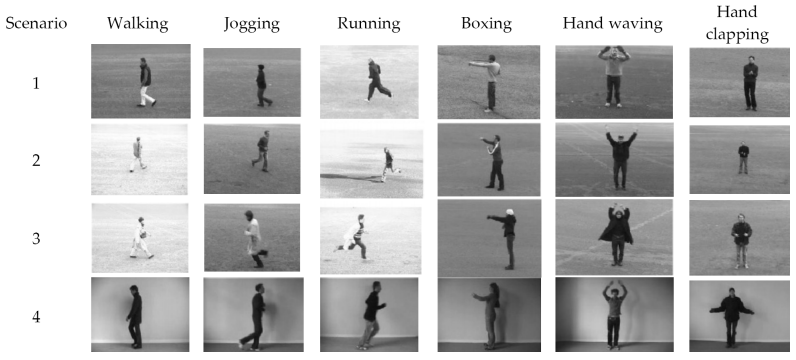


FIGURE 2.12: One frame example of each action from four different scenarios in the KTH dataset

a top camera. The extracted silhouettes of the video sequences are also provided for experimentation. Basically, two types of methods have been proposed for multiview action recognition, i.e., 2D and 3D-based methods. The 3D based methods have reported higher accuracy than the 2D based methods on this dataset but at a higher computational cost. The highest accuracy reported on this dataset is 100% by Holte et al [143] using 3D motion descriptors (HOF3D descriptors and 3D spatial pyramids (SP)). The example frames for each action from five different camera views are shown in Figure 2.13.

2.4.4 HMDB-51

The HMDB-51 [72] is one of the largest datasets available for activity recognition developed by Serre lab, Brown University, USA in 2011. It consists of 51 types of daily life actions comprised of 6849 video clips collected from different sources such as movies, YouTube, and Google videos. The highest accuracy reported so far on this dataset is 74.7% by Yu et al [181] using SP-CNN as shown in Table 2.4. One frame example for each action is shown in Figure 2.14 and Figure 2.15.

2.4.5 Hollywood2

Hollywood2 [200] action dataset was created by INRIA (Institut National de Recherche en Informatique et en Automatique), France in 2009. This dataset consists of 12 actions (get out of car, answer phone, kiss, hug, handshake, sit down, stand up, sit up, run, eat, fight, and drive car) with dynamic background features. This dataset is very challenging, consists of short unconstrained movies with multiple persons, cluttered background, camera motion, and large intra-class variations. This dataset is meant for evaluation of HAR algorithms in real life scenarios. Many researchers have evaluated their algorithms on

FIGURE 2.13: One frame example for each action from five different camera views in IXMAS dataset

this dataset, the best accuracy achieved so far is 75.2% by Fernando et al [193] using rank pooling and CNN. Some example frames from Hollywood2 dataset are shown in Figure 2.16.

FIGURE 2.14: Exemplar frames for action 1 to 28 from HMDB-51 action dataset

FIGURE 2.15: Exemplar frames for action 29 to 51 from HMDB-51 action dataset

FIGURE 2.16: Exemplar frames from Hollywood2 dataset

TABLE 2.4: Well-known public datasets for human activity recognition

Dataset	Year	No. of Actions	Method	Accuracy (%)
KTH	2004	6	[69]	98.2
Weizmann	2005	9	[112]	100
IXMAS	2006	13	[143]	100
UCF Sports	2008	10	[69]	95.0
Hollywood2	2009	12	[193]	75.2
YouTube	2009	11	[77]	93.38
HDMB-51	2011	51	[181]	74.7
UCF-101	2012	101	[181]	91.6
ActivityNet (Untrimmed)	2015	200	[199]	42.2
ActivityNet (Trimmed)	2015	200	[199]	50.2

2.4.6 UCF-101 Action Recognition Dataset

UCF-101 action recognition dataset [188] was created by the Centre for Research in Computer Vision, University of Central Florida, USA in 2012. This is one of the largest action dataset contains 101 action categories collected from YouTube. This dataset is an extension of UCF-50 [201] dataset with 50 action categories. UCF-101 contains 13,320 videos in total, aimed at encouraging the researchers to develop their algorithms for human action recognition in realistic scenarios. The example frames for each action are shown in Figure 2.17 and Figure 2.18.

2.4.7 UCF Sports Action Dataset

UCF sports action dataset was created by the Centre for Research in Computer Vision, University of Central Florida, USA in 2008 [16, 17]. It consists of 11 sports action categories (walking, swing-side, swing-bench, skateboarding, running, lifting, kicking, golf swing, riding, and diving) broadcasted on television channels. The dataset includes total 150 video sequences of realistic scenarios. The best accuracy achieved on this dataset so

FIGURE 2.17: Exemplar frames for actions 1 to 57 from UCF-101 dataset

FIGURE 2.18: Exemplar frames for actions 58 to 101 from UCF-101 dataset

far is 95.0% by Sadanand and Corso [69] using STVs as shown in Table 2.4. The example frames for each action are shown in Figure 2.19

2.4.8 YouTube Action Dataset

YouTube action dataset [89] was developed in 2009. This is a challenging dataset due to camera motion, viewpoint variations, illumination conditions, and cluttered backgrounds. It contains 11 action categories: biking, diving, basketball shooting, horse riding, swinging, soccer juggling, trampoline jumping, volleyball spiking, golf swinging, tennis swinging, and walking with a dog. The highest accuracy achieved so far on this dataset is 93.38% by Peng et al [77] using FV and SFV. The example frames for each action are shown in Figure 2.20.

2.4.9 ActivityNet Dataset

ActivityNet [199] was created in 2015. This is a large-scale video dataset covering wide range of complex human activities. It provides 203 action categories in total 849 hours of video data. This dataset is specifically helpful for training the classifiers which require a huge amount of data for training such as deep neural networks. According to the results reported by Heilbron et al [199], authors achieved 42.2% accuracy on untrimmed

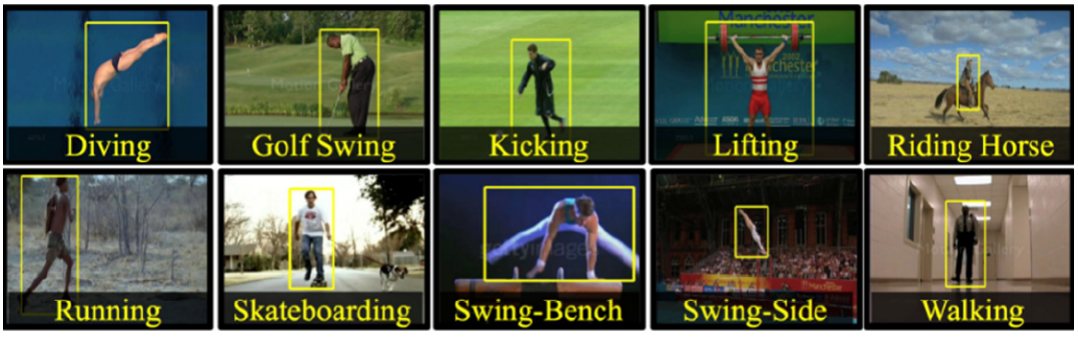


FIGURE 2.19: Exemplar frames from sports action dataset

FIGURE 2.20: Exemplar frames of 11 sports actions from YouTube action dataset

videos and 50.2% on trimmed videos classification. They used deep features (DF), motion features (MF), and static features (SF) as shown in Table 2.3. Some example frames from this dataset are shown in Figure 2.21.

2.5 Conclusions

This chapter provided a comprehensive literature review of state-of-the-art human activity representation and recognition approaches including both handcrafted and learning-based representations. The handcrafted activity representation approaches have been there for a quite long time. These approaches have achieved good results on different

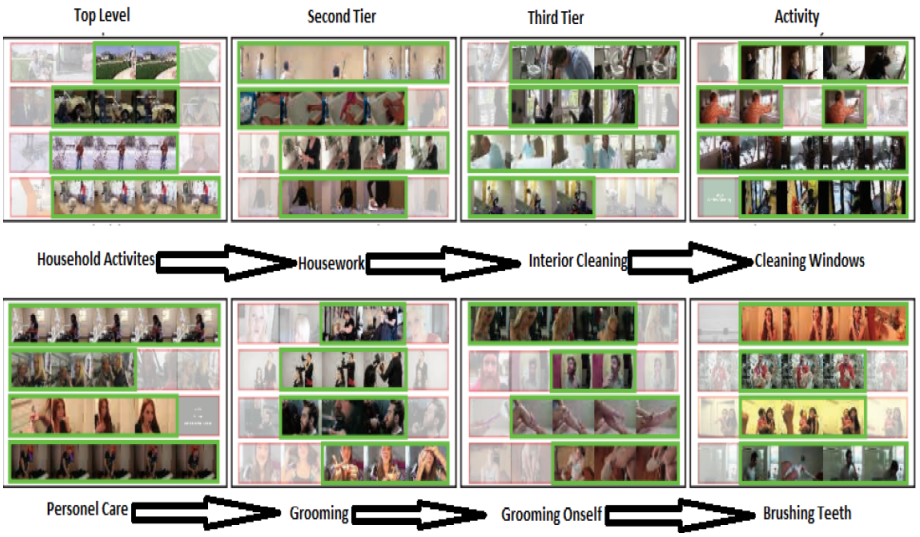


FIGURE 2.21: Exemplar frames from ActivityNet dataset

publically available benchmark datasets. However, most successful handcrafted representation methods are based on the local densely-sampled descriptors, which incur a high computational cost. In these approaches, the important features from the sequence of image frames are extracted to build the feature vector using human engineered feature detectors and descriptors. Then, the classification is performed by training a generic classifier. These approaches include space-time, appearance based, structural feature-base, local binary patterns, and fuzzy logic-based approaches. However, handcrafted representation-based approaches are still widely used due to computational complexity and dataset requirement of deep learning techniques for activity recognition.

On the other hand, learning-based action representation approaches, specifically, deep learning based techniques use trainable feature extractors followed by a trainable classifier, which lead to the concept of end-to-end learning or learning from pixel level to action categories identification.This eliminates the need for handcrafted feature detectors and

descriptors used for action representation. These approaches include evolutionary, dictionary learning, and deep learning-based approaches. Recently, the research community has paid a lot of attention to these approaches. This is mainly due to their high performance as compared to their handcrafted counterparts on challenging datasets. However, fully data-driven deep models referred to as black-box have some limitations: Firstly, it is difficult to incorporate problem-specific prior knowledge into these models. Secondly, some of the best performing deep learning-based methods are still dependent on handcrafted features. In HAR, the performance of the pure learning-based methods is still not up to the mark. This is mainly due to the unavailability of huge datasets for action recognition unlike in the object recognition where huge dataset such as ImageNet is available.

In order to provide further insight into the field, some well-known public datasets for activity recognition are presented for experimentations and evaluations of HAR techniques. These datasets include: KTH, Weizmann, IXMAS, UCF Sports, Hollywood2, YouTube, HMDB-51, UCF-101, and ActivityNet.

Chapter 3

Human Action Recognition From Multiple Views

IN THIS CHAPTER

A novel method for human action recognition from multiple views, i.e., view invariant action recognition is presented. The introduction of the proposed method is presented in Section 3.1. While, state-of-the-art view invariant action recognition techniques are ascertained in Section 3.2. The proposed methodology and experimentations results are explained in Section 3.3 and 3.4 respectively. Finally, the chapter is concluded in Section 3.5.

3.1 Introduction

Over the past decade, many techniques have been proposed for human action recognition. However, it is still a challenging task due to many issues involved in it [30]. The major challenges and issues in HAR are as follows: (1) occlusion; (2) variation in human

appearance, shape, and clothes; (3) cluttered backgrounds; (4) stationary or moving cameras; (5) different illumination conditions; and (6) viewpoint variations. Among these challenges, viewpoint variation is one of the major problems in HAR since most of the approaches for human activity classification are view-dependent and can recognize the activity from one fixed view captured by a single camera. These approaches supposed to have the same camera view during training and testing. This condition cannot be maintained in real world application scenarios. Moreover, if this condition is not met, their accuracy decreases drastically because the same actions look quite different when captured from different viewpoints [3, 202]. A single camera-based approach also fails to recognize the action when an actor is occluded by an object or when some parts of the action are hidden due to unavoidable self-occlusion. To avoid these issues and get the complete picture of an action, more than one camera is used to capture the action, this is known as action recognition from multiple views or view-invariant action recognition [203].

There are two major approaches for recognition of human actions from multiviews: 3D approach and 2D approach [5]. In the first approach, a 3D model of a human body is constructed from multiple views and motion representation is formed from it for action recognition. This model can be based on cylinders, ellipsoids, visual hulls generated from silhouettes, or surface mesh. Some examples of motion representation are 3D optical flow [106], shape histogram [107], motion history volume [108], 3D body skeleton [109], and spatio-temporal motion patterns [204]. Usually, the 3D approach provides higher accuracy than a 2D approach but at higher computational cost, which makes it less applicable

for real time applications. In addition to this, it is difficult to reconstruct a good-quality 3D model because it depends on the quality of extracted features or silhouettes of different views. Hence, the model is exposed to deficiencies which might have occurred due to segmentation errors in each viewpoint. Moreover, a good 3D model of different views can only be constructed when the views overlap. Therefore, a sufficient number of viewpoints have to be available to reconstruct a 3D model.

However, recently some 3D cameras have been introduced for capturing images in 3D form. Among these, 3D time-of-flight (ToF) cameras and Microsoft Kinect have become very popular for 3D imaging. These devices overcome the difficulties that are faced by the classical 3D multiview action-recognition approaches when reconstructing a 3D model. However, these sensors also have several limitations. For example, in contrast to a fully reconstructed 3D model from multiple views, these sensors only capture the frontal surfaces of the human and other objects in the scene. In addition to this, these sensors also have limited range about 67 m, and data can be distorted by scattered light from the reflective surfaces [51]. Due to the limitations of the 3D approach, researchers prefer to employ a 2D approach for human-action recognition from multiple views [205].

The methods based on 2D models extract features from 2D images covering multiple views. Different methods have been proposed for multiview action recognition based on 2D models. However, three important lines of work are mentioned here. The first approach handles it at feature level, achieves view-invariant action representation using appropriate feature descriptor(s) or fusion of different features [130, 206], and then action recognition is performed using an appropriate classifier. The second one handles it at a

classification level by determining the appropriate classification scheme. The classification is carried out either by a single universal classifier or multiple classifiers are trained, and later on their results are fused to get the final result [207, 208]. The third one utilizes the learning-based model, such as deep learning and its variations, to learn the effective and discriminative features directly from the raw data for multiview action recognition [189, 209].

The proposed method in this chapter falls under first category. This method is based on a novel feature descriptor for multiview human action recognition. This descriptor employs region-based features extracted from the human silhouette. To achieve this, the human silhouette is divided into regions in a radial fashion with the interval of a certain degree, and then region-based geometrical and Hu-moments features are obtained from each radial bin to articulate the feature descriptor. A multiclass support vector machine classier is used for action classication. The proposed approach is quite simple and achieves state-of-the-art results without compromising the efciency of the recognition process.The contribution is two-fold. Firstly, the proposed approach achieves high recognition accuracy with simple silhouette-based representation. Secondly, by taking advantage of this simple representation, the proposed provide high recognition speed of 34 frames per second on a challenging multiview IXMAS (INRIA Xmas Motion Acquisition Sequences) dataset. Due to its high accuracy and speed the proposed approach is suitable for real-time applications.

3.2 Related Work

This section presents state-of-the-art methods for multiview action recognition based on a 2D approach. These methods extract features from 2D image frames of all available views and combine these features for action recognition. Then, classifier is trained using all these viewpoints. After training the classifier, some methods use all viewpoints for classification [210], while others use a single viewpoint for classification of a query action [211–213]. In both cases, the query view is part of the training data. However, if the query view is different than the learned views, this is known as cross-view action recognition. This is even more challenging than the multiview action recognition [150, 214].

Different types of features such as motion, shape, or combination of motion and shape features have been used for multiview action recognition. In [212], silhouette-based features were acquired from five synchronized and calibrated cameras. The action recognition from multiple views was performed by computing the R transform of the silhouette surfaces and manifold learning. In [114], contour points of the human silhouette were used for pose representation, and multiview action recognition was achieved by the arrangements of multiview key poses. Another silhouette-based method was proposed in [115] for action recognition from multiple views; this method used contour points of the silhouette and radial scheme for pose representation. Then, model fusion of multiple camera streams was used to build the bag of key poses, which worked as a dictionary for known poses and helped to convert training sequences into key poses for a sequence-matching algorithm. In [130], a view-invariant recognition method was proposed, which

extracted the uniform rotation-invariant local binary patterns (LBP) and contour-based pose features from the silhouette. The classification was performed using a multiclass support vector machine. In [117], scale-invariant features were extracted from the silhouette and clustered to build the key poses. Finally, the classification was done using a weighted voting scheme.

Optical flow and silhouette-based features were used for view-invariant action recognition in [120], and principal component analysis (PCA) was used for reducing the dimensionality of the data. In [122], coarse silhouette features, radial grid-based features and motion features were used for multiview action recognition. Another method for viewpoint changes and occlusion-handling was proposed in [211]. This method used the histogram of oriented gradients (HOG) features with local partitioning, and obtained the final results by fusing the results of the local classifiers. A novel motion descriptor based on motion direction and histogram of motion intensity was proposed in [118] for multiview action recognition followed by a support vector machine used as a classifier. Another method based on 2D motion templates, motion history images, and histogram of oriented gradients was proposed in [119]. A hybrid model which combines convolution neural networks (CNN) with hidden Markov model (HMM) was used for action classification [209]. In this method, the CNN was used to learn the effective and robust features directly from the raw data, and HMM was used to learn the statistical dependencies over the contiguous actions and conclude the action sequences.

3.3 Proposed System

In recent years, various methods have been published for multiview action recognition, but very few are actually suitable for real-time applications due to their high computational cost. Therefore, the cost of the feature extraction and action classification has to be reduced as much as possible. The proposed system has been designed around these parameters. It uses the human silhouette as input for feature extraction and formulating the region-based novel feature descriptor. Human silhouette can easily be extracted using foreground detection techniques. At the moment our focus is not on foreground segmentation; rather, we are focused on later phases of the action recognition such as feature extraction and classification. Although the human silhouette is a representation that contains much less information than the original image, we show that it contains sufficient information to recognize the human actions with high accuracy. This work proposes a novel feature descriptor for multiview human-action recognition based on human silhouette. When a silhouette is partitioned into radial bins, then region descriptors are formed which are represented as a triangle or quadrangle. Accordingly, it makes the pose description process simple and accurate. The descriptive detail of the proposed system is given in Algorithm 1 and block diagram is shown in Figure 3.1.

This work considers the centeroid of a silhouette as a point of origin, and divide the human silhouette into R radial bins with the angular interval of the same degree. Here the value of the R is selected to be 12, which constitutes 12 radial bins with the angular interval of 30°each. This value has been selected with experimentation to cover the maximum

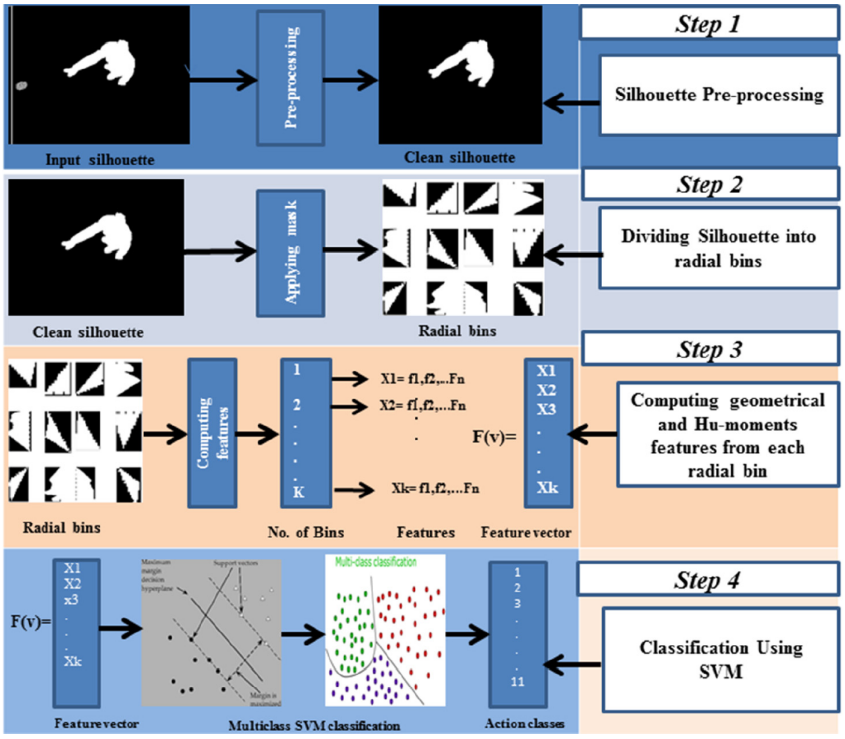


FIGURE 3.1: Block diagram of the proposed multiview HAR system

Algorithm 3.1 Feature Extraction and Classification for HAR from multiple views

1. Input video from multiple cameras
2. Extraction of silhouettes from the captured video
3. Feature Extractions:
   - Division of the human silhouette into radial bins
   - Computation of region-based geometrical features from each radial bin
   - Computation of Hu-moments features from each radial bin
4. Action classifications by multiclass support vector machine
5. Output recognized action

viewing angles. This is unlike [215], where the radial histograms were used as a feature descriptor [115], and the contour points of each radial bin were considered as features, and model fusion was used to achieve the multiview action recognition by obtaining key poses for each action through K-means clustering. The proposed method computes efficient and discriminative geometrical and Hu-moment features from each radial bin of the silhouette itself. Then, these features from all bins are concatenated into a feature vector for multiview action recognition using support vector machine.

This approach has three major advantages. Firstly, it divides the silhouettes into radial bins, which cover almost all viewing angles, thus provides an easy way to compute the features for different views of an action. Secondly, it uses the selected discriminative features and avoids extra computation such as fusion or clustering as used in [115]. Thirdly, due to employing selected features from each bin, it does not require any dimensionality reduction technique.

3.3.1 Pre-Processing

Usually, some lines or small size regions are formed around the silhouette due to segmentation errors, loose clothing of the subject under consideration, and other noise. These unnecessary regions do not offer any important information for action recognition, but rather create problems for the feature extraction algorithm and increase the complexity. By removing these small regions, complexity can be reduced and the feature-extraction process can be made more accurate. In our case, a region was considered to be small and unnecessary if its area is less than 1/10 of the silhouette. This threshold was set based on an observation on 1000 silhouette images of different actors and actions. An example of

the silhouette before and after noise removal is shown in first row of Figure 3.1.

3.3.2 Multiview Features Extraction and Representation

The success of any recognition system mainly depends on proper feature selection and extraction mechanism. For action recognition from different views, a set of discriminative and view-invariant features have to be extracted.The proposed feature descriptor is based on two types of features: (1) region-based geometric features and (2) Hu-moments features extracted from each radial bin of the silhouette. The overview of the feature extraction process is shown in Figure 3.2.

$$C_m = (x_{c,}\, y_c) \tag{3.1}$$

where

$$x_c = \frac{\sum_{i=1}^{n} x_i}{n} \qquad and \qquad y_c = \frac{\sum_{i=1}^{n} y_i}{n} \tag{3.2}$$

3.3.2.1 Region-Based Geometric Features

The shape of a region can be described by two types of features. One is the region-based and the other is boundary-based features. The region-based features are less affected by noise and occlusion than boundary-based features. Therefore, region-based features are better choice to describe the shape of a region [111]. Generally, a shape is described by a set of vectors known as descriptors. A good descriptor is one which has the ability to reconstruct the shape from the feature points. The resultant shape should be an approximation of the original shape and should yield similar feature values. These features have

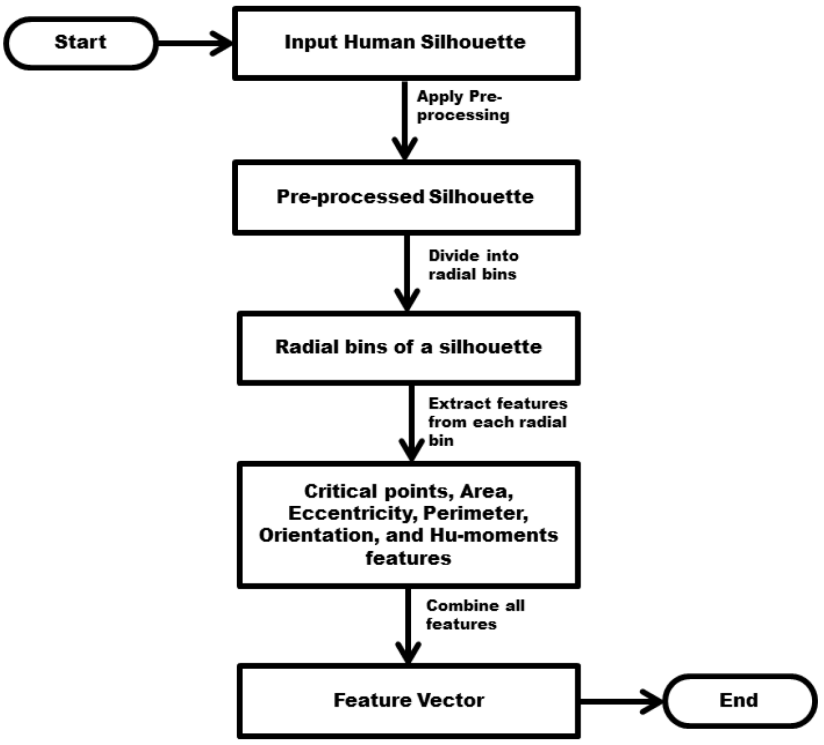


FIGURE 3.2: Overview of feature extraction process

been computed as follows:

1. First of all, the centroid of a human silhouette is calculated using Equation 3.2, as shown in Figure 3.3

2. After computing the centroid by Equation 3.2, a radius of the silhouette is computed.

3. The silhouette is divided into 12 radial bins with respect to its centroid using a mask; this division has been made with intervals of 30°. These bins are shown in Figure 3.4.

4. The following region-based features are computed for each bin of the silhouette:

- **Critical points:** As we move the mask on the silhouette in counter-clockwise direction, a triangle or quadrangle shape is formed in each bin. We compute the critical points (corner points) of each shape and their distances. There can be different numbers of critical points for each shape; therefore, the mean and variance of these points have been computed as features. It gives us $2 \times 12 = 24$ features for each silhouette.
- **Area:** The simple and natural property of a region is its area. In the case of a binary image, it is a measure of size of its foreground. We have computed the area of each bin, which provides 12 important features for each human silhouette.
- **Eccentricity:** The ratio of the major and minor axes of an object is known as eccentricity. In our case, it is ratio of distance between the major axis and foci of the ellipse. Its value is between 0 and 1, depending upon the shape of the ellipse. If its value is 0, then actually it is a circle; if its value is 1, then it is a line segment. We have computed eccentricity for each of the 12 bins.
- **Perimeter:** This is also an important property of a region. The distance around the boundary of a region can be measured by computing the distance between each pair of pixels. We have computed perimeter of each bin forming a triangle or quadrangle.
- **Orientation:** It is an important property of a region which specifies the angle between x-axis and major axis of the ellipse. Its value can be between $-90°$ and $+90°$

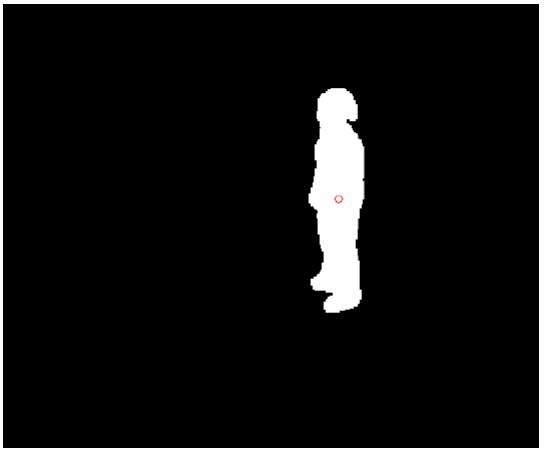


FIGURE 3.3: Example image of the human silhouette with centroid

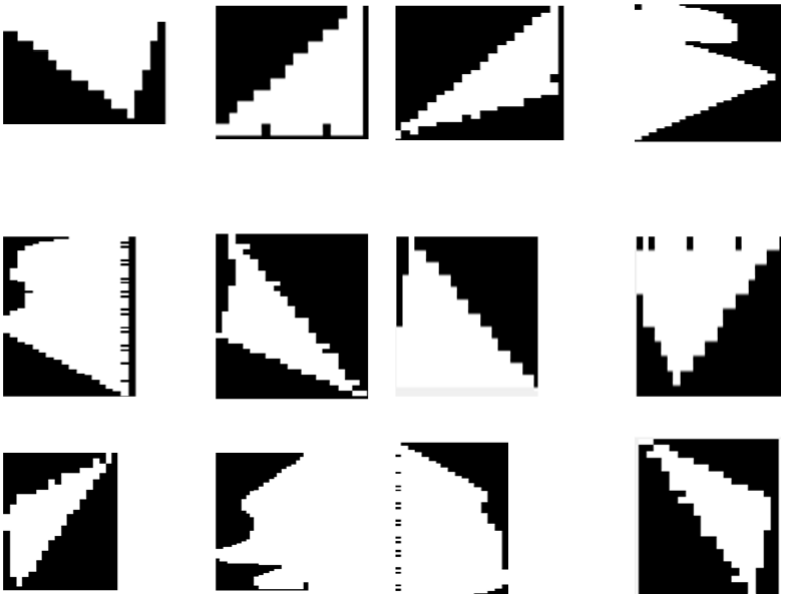


FIGURE 3.4: Example of division of silhouette into radial bins

3.3.2.2 Hu-Moments Invariant Features

The use of invariant moments for binary shape representation was proposed in [216]. The moments which are in-variants with respect to rotation, scales, and translations, are known as Hu-moments in-variants. The proposed method computes seven moments for each radial bin of the silhouette. The moment *(p,q)* of an image f(x,y) of size $M \times$N is

defined as:

$$m_{p,q} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y)\, x^p y^q \tag{3.3}$$

Here p is order of x and q is order of y.The central moments can be calculated in the same way as these moments except the values of x and y is displaced by the mean values as follows:

$$\mu_{p,q} = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} f(x,y)\, (x - x_{avg})^p (y - y_{avg})^q \tag{3.4}$$

where

$$x_{avg} = \frac{m_{10}}{m_{00}} \quad and \quad y_{avg} = \frac{m_{01}}{m_{00}} \tag{3.5}$$

By applying normalization, scale invariant moments are obtained. Hence, normalized central moments are defined as follows [217].

$$\eta_{p,q} = \frac{\mu_{p,q}}{\mu_{00}^{\gamma}}, \gamma = \frac{p+q+2}{2}, p+q = 2,3,... \tag{3.6}$$

Based on these central moments, [216] introduced seven Hu-moments as linear combination of central moments defined as follows:

$$h1 = \eta_{20} + \eta_{02} \tag{3.7}$$

$$h2 = (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \tag{3.8}$$

$$h3 = (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \tag{3.9}$$

$$h4 = (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \tag{3.10}$$

$$\begin{aligned} h5 = {} & (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ & + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \end{aligned} \tag{3.11}$$

$$h6 = (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \tag{3.12}$$

$$\begin{aligned} h7 = {} & (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\ & - (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \end{aligned} \tag{3.13}$$

3.3.3 Action Classification with SVM Multiclass Classifier

After computing the features from video, an appropriate classifier has to be used for classification of the actions. In supervised learning, discriminative models are more effective than generative models [218]. Support vector machine (SVM) is one of the most successful classifiers in discriminative models category. It has shown better performance than some old-style classifiers such as backpropagation neural networks, Naïve Bayes, and k-nearest neighbors (KNNs) in many classification problems [62]. The SVM was first proposed in [160], originally, it was developed for binary classification but later on extended to the multiclass classification problem. There are two main approaches for multiclass SVM. The first approach considers all classes of data directly into one optimization formulation, while the second approach constructs and combines binary classifiers to build a multiclass classifier. The second approach is computationally less expensive and easy to implement. Many algorithms have been derived for multiclass classification using this approach, such as one-against-all [219], one-against-one [220], Directed Acyclic

Graph-Support Vector Machine (DAG-SVM) [221], Error-Correcting Output Codes Support Vector Machine (ECOC-SVM) [222], and Support Vector Machines with Binary Tree Architecture (SVM-BTA) [223]. Among these, one-against-all [219] and one-against-one [220] are two commonly used methods for multiclass classification. The one-against-all needs N SVM binary classifiers for an N class classification problem, while one-against-one method needs (NxN-1)/2 binary classifiers for an N number of classes, each trained from samples of two corresponding classes. As compared to one-against-all method, one-against-one is better in terms of accuracy for many classification problems [224].

The proposed method used multiclass SVM classifier implementation in [54] for multi-view action recognition, which uses the one-against-one method with radial basis function (RBF) kernel. Moreover, to estimate the best parameters for classifier, grid search was conducted to know the best value for parameter γ and C. Here, γ represents the width of the RBF kernel and C represents the weight of error penalty. The appropriate set of (C,γ) increases the overall accuracy of the SVM classifier [225].

3.4 Experimentations

For the evaluation of proposed method, comprehensive experimentations have been conducted on a well-known multiview IXMAS [141] dataset. The leave-one-sequence-out (LOSO) scheme has been used for view-invariance evaluation. In this scheme, the classifier is trained on all sequences except one, which is used for testing. This process is repeated for all possible combinations and results are averaged. This is a common strategy used by different researchers such as [118, 211] for evaluation of their methods.

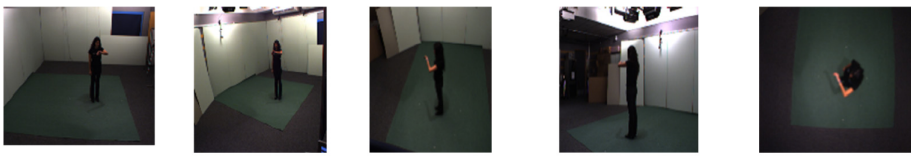
This is helpful to compare the proposed method with state-of-the-art methods.

3.4.1 Evaluation on Multiview Action Recognition Dataset

The IXMAS is a challenging and well-known dataset with multiple actors and camera views. This dataset is popular among the human-action recognition methods for testing view-invariant action recognition algorithms, including both multiview and cross-view action recognition. It includes 13 daily life action classes with 5 different cameras, including one top-view and four side cameras, as shown in Figure 3.5. Each action is performed 3 times by 12 different subjects while actors keep changing orientations in each sequence during action execution. The change in orientation is indicated by action labels, and no additional information is provided other than these labels. Most of the existing methods use selected action classes and actors for experimentation [118, 211]. For comparison, this work uses 11 action classes performed by 12 actors as shown in Figure 3.6. The name and the label index of these actions are shown in Table 3.1.

TABLE 3.1: Action class names used in experimentation

Index	**Action name**	**Index**	**Action name**
1	Checking watch	7	Walking
2	Crossing arms	8	Waving
3	Scratching head	9	Punching
4	Sitting down	10	Kicking
5	Getting up	11	Picking up
6	Turning around		



Camera 1 *Camera 2* *Camera 3* *Camera 4* *Camera 5*

FIGURE 3.5: Five cameras views of check watch action from IXMAS dataset

3.4.2 Comparison with Similar Methods on IXMAS Dataset

The proposed method achieved a recognition rate of 89.75% with leave-one-sequence-out (LOSO) cross-validation on 11 actions. The recognition rate of individual action is presented in a confusion matrix, shown in Figure 3.7. The results confirm that proposed method outperforms the similar state-of-the-art 2D based methods such as [63, 114, 118, 206, 211, 213, 226–229] recorded in Table 3.2. It is important to be mentioned here that the number of classes, actors, and views used in experimentations vary among these methods. For example, in [63], 89.4% accuracy has been reported but they excluded camera 4 from experimentation. Likewise, [229] also excluded the top camera and considered only the remaining 4 cameras. Moreover, most of the published methods are not appropriate for real-time application due to their high computational cost. The proposed method considers all views of the IXMAS dataset including the top view for recognition. The results indicate that the proposed method is superior to the similar 2D methods, not only in recognition accuracy but also in recognition speed as well.

TABLE 3.2: Comparison with state-of-the-art methods on IXMAS dataset

Year	Method	Accuracy (%)
2016	**Proposed method**	**89.75**
2016	Chun et al. [118]	83.03
2013	Chaaraoui et al. [114]	85.9
2011	Wu et al. [63]	89.4
2011	Junejo et al. [206]	74
2010	Weinland et al.[211]	83.4
2009	Reddy et al. [226]	72.6
2008	Liu and Shah [213]	82.8
2008	Cherla et al. [229]	80.1
2008	Vitaladevuni et al. [228]	87.0
2007	Lv and Nevatia [227]	80.6

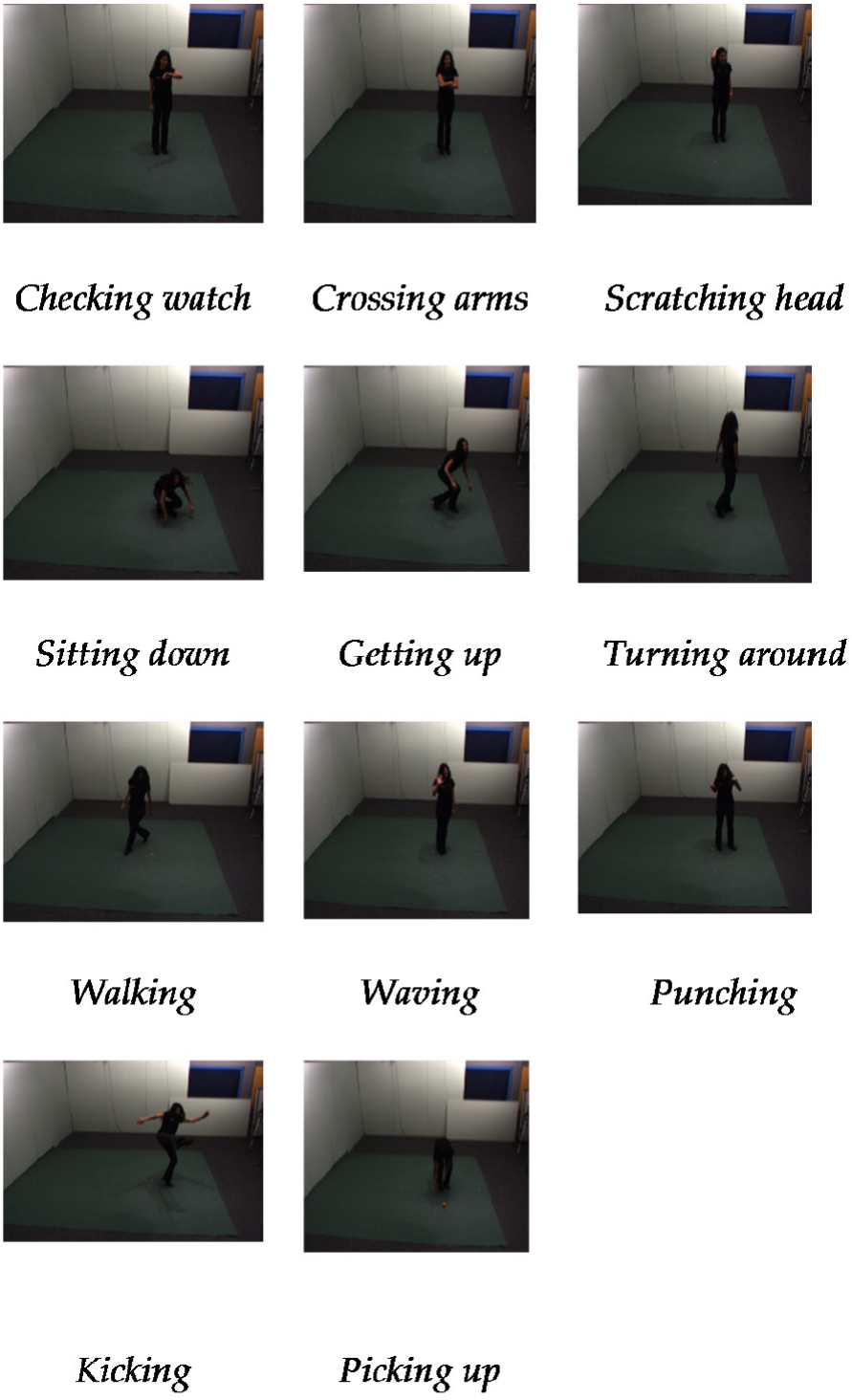


FIGURE 3.6: One frame example of 11 actions in IXMAS dataset

TABLE 3.3: Comparison of average testing speed on IXMAS dataset

Method	**Average FPS**
Proposed method	**34**
Chaaraoui et al. [114]	26
Cherla et al. [229]	20
Lv and Nevatia [227]	5.1

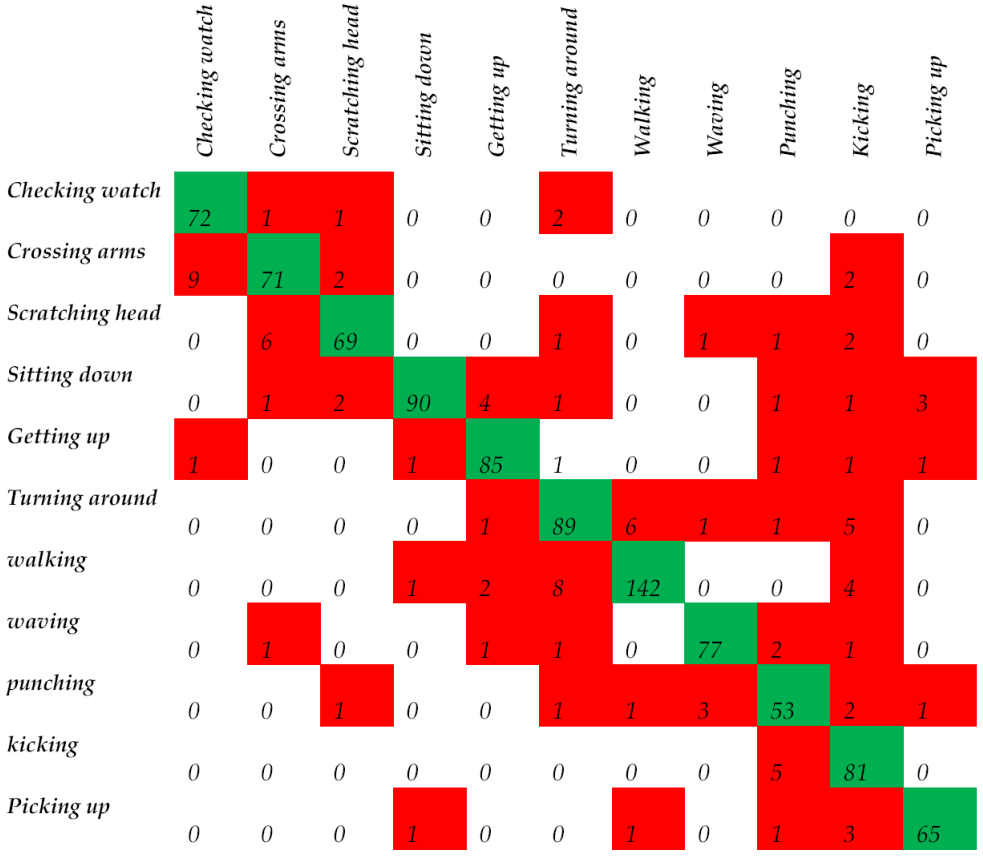


FIGURE 3.7: Confusion matrix of IXMAS dataset with 11 actions

The resolution of IXMAS dataset is only 390×291, which is very low resolution as compared to many other action recognition datasets. For experimentation MATLAB R2015b implementation on Intel Core i7-4770 CPU with 8 cores @ 3.4 GHz, 8 GB RAM, and Windows 10 operating system were used. However, only 4 cores of the CPU were utilized during the experimentation. Moreover, no optimization techniques were used in the code. The average testing speed of the proposed method is 2.88 ms per frame, which

correspond to almost 34 frames per second (FPS); this is much better than the existing methods as recorded in Table 3.3

3.5 Conclusions

In this chapter, a multiview human-action recognition method based on a novel region-based feature descriptor is presented. The multiview HAR methods can be divided into two categories: 2D approach and 3D approach-based methods. Usually, the 3D approach provides better accuracy than a 2D approach, but it is computationally expensive, which makes it less applicable for real-time applications. The proposed method uses the human silhouette as input to the features extraction process. Although the human silhouette contains much less information than the original image, but proposed method practically confirms that it is sufficient for action recognition with high accuracy. Moreover, to get the view-invariant features, the silhouette is divided into radial bins with the angular interval of 30°each. Then, carefully selected region-based geometrical features and Hu-moment features are computed for each bin. For action recognition, a multiclass support vector machine with RBF kernel was employed. The proposed method was evaluated on the well-known IXMAS dataset. This dataset is a challenging benchmark available for evaluation of multiview action recognition methods. The results indicate that the proposed method outperforms the state-of-the-art 2D methods both in terms of speed and recognition accuracy. The testing speed of the proposed method is 34 frames per second, which makes it very much suitable for real-time applications. As far as more complex datasets such as HMDB-51, YouTube, and Hollywood-II are concerned, the proposed method was

not tested on these datasets. However, authors believe that proposed method should also produce good accuracy with these datasets as well, because once the silhouette is extracted the scene complexities should not matter. However, perfect silhouette extraction in complex scenarios is still a challenging task which may further affect the feature extraction process as well. As a future work authors would like to extend this work for cross-view action recognition, a special case of multiview action recognition, where a query view is different than the learned views with more complex dataset.

Chapter 4

Human Action Recognition Using Transfer Learning

IN THIS CHAPTER

An innovative method for human action recognition using transfer learning is presented. Transfer learning is defined as using a pre-trained model on specific dataset for solving a new task, even in a different domain. The introduction of the proposed method is presented in Section 4.1, and related work is presented in Section 4.2. The proposed methodology and experimentations results are explained in Section 4.3 and 4.4 respectively. Finally, the chapter is concluded in Section 4.5.

4.1 Introduction

In recent years, learning-based representation, and in particular deep learning, has introduced the concept of end-to-end learning by using the trainable feature extractor followed by a trainable classifier [19, 146]. These approaches have revealed remarkable progress for action recognition in videos. The deep learning model, introduced in

[162] for reducing the dimensionality of the data, CNN [230] and Deep Belief Networks (DBNs) [161] have been widely used for image classification, object recognition, and action recognition.

However, training a new deep learning model from scratch requires huge amount of data, high computational resources, and hours, in some cases days, of training. In real-world applications, collecting and annotating a huge amount of domain-specific data is time consuming and expensive. Hence, collecting the sufficient amount of data may not be a viable option in many cases [231, 232], which makes it quite challenging to apply deep learning models for such problems. For combating this challenge, researchers revisited their strategies for visual categorization to make them in-line with the working of the human vision system. Humans have capability to learn thousands of categories in their lives from just few samples. It is believed that humans achieve this capability by accumulating knowledge over a time period and transferring it for learning new objects [233]. Researchers are convinced that the knowledge of previous objects assist in the learning of new objects through their similarity and connection with the new objects. Based on this idea, some studies suggest that the deep learning models trained for a classification task, can be employed for a new classification task [234–236]. Thus, the CNN models trained on a specific dataset or task can be fine-tuned for a new task even in a different domain [185, 237, 238]. This concept is known as transfer learning or domain adaptation.

Transfer learning has been studied as a machine learning technique for a long time, for solving different visual categorization problems. In recent years, due to an explosion of information such as images, audio, and video over the internet, demands for high

accuracy, and computational efficiency have increased. Due to these reasons, transfer learning has attracted a lot of interest in the areas of machine learning and computer vision. When traditional machine learning techniques have reached their limits, transfer learning unlocks new flow streams for visual categorization. It has primarily changed the approach, the way machines used to learn and treat the classification tasks [239].

Transfer learning mainly employs two approaches: 1) preserving the original pre-trained network and updating the weights based on the new training dataset, 2) using the pre-trained network for feature extraction and representation, followed by a generic classifier such as SVM for classification [170]. The second approach has been successfully applied for many recognition and classification tasks [234, 240]. The proposed technique for human action recognition also falls under the second category. In this work, the recently proposed benchmark for deep models such as AlexNet [168], and GoogleNet [241] are investigated. Based on experimentation, the AlexNet has been selected as source model for building a target model. The source model has been used for feature extraction and representation followed by a hybrid (SVM-KNN) classifier for action recognition.

4.2 Related Work

This section discusses the literature review on the existing methods for action recognition using handcrafted representations and deep learning. Action recognition using handcrafted feature descriptors, such as extended SURF[242], HOG-3D [243], and some other shape and motion based features descriptors [20, 73, 78, 244, 245], have achieved good performance for human action recognition. However, these approaches have several

limitations: In particular, handcrafted feature-based techniques require expert designed feature detectors, descriptors, and vocabulary building methods for feature extraction and representation, this feature engineering process is labor-intensive and requires expertise of the subject matter.

Due to these limitations, more research is directed to a deep learning-based approach. This approach has been used in several domains such as image classification, speech recognition, and object recognition, just to name few [246]. These models have also been explored for human activity recognition. Some prominent contributions like 3D ConvNets [173], Convolutional RBMs [247], learning spatio-temporal with 3D ConvNets [179], Deep ConvNets [172], and Two-stream ConvNets [2], have achieved good results. The on-line deep learning method such as [18], is also getting more attention for action recognition. In [248], a human action recognition method was proposed using unsupervised on-line deep learning technique. This method achieved accuracy of 89.86%, and 88.5% on KTH and UCF sports dataset, respectively.

Of the handcrafted feature-based techniques, trajectory based methods in particular, have less discriminative power. Conversely, deep network architectures are inefficient in capturing the salient motion. For addressing this issue, [189] combined deep convolutional networks with trajectory based features for recognizing human actions. However, deep learning-based methods also have some limitations. These models require huge dataset for training, and collecting huge amount of domain-specific data is time consuming and expensive. Therefore, training the deep learning model from scratch is not feasible for domain-specific problems. This problem can be solved using a pre-trained

network as a source architecture for training the target model with small dataset, known as transfer learning [170].

Fortunately, the winner models of ImageNet Large Scale Visual Recognition Challenge (ILSVRC) such as AlexNet [168], GoogleNet [241], and ResNet [249] are publicly available as pre-trained networks. These networks can be used for transfer learning. One of the important ways to employ the existing models for a new task is to use pre-trained models as a feature extraction machine and combine this deep representation with off-the-shelf classifiers for action recognition [234]. Some researchers have also used cross-domain knowledge transfer for action recognition. In [250], a cross-domain knowledge transfer was performed between the KTH, TRECVID [251] and Microsoft research action datasets. The TRECVID and Microsoft research action datasets were used as a source domain, while KTH was used a target domain. In addition to this, some researchers have used cross-view knowledge transfer, which is a special form of cross-domain knowledge transfer for multiview action recognition.

4.3 Methodology

In machine learning, utilizing the previously learned knowledge for solving a new task is known as transfer learning or knowledge transfer [252]. The transfer learning using deep CNNs is very helpful for training the model with a limited size dataset, because CNNs are prone to over-fitting with small datasets. Over-fitting can be avoided by increasing the size of the training data, however, it is very difficult and expensive to provide the large amount of annotated data. In this situation, transfer learning comes in handy

and solves this problem by using the pre-trained deep representation as a source architecture for building the new architecture [253]. In this work, the AlexNet [168] has been employed as a source architecture for solving the human action recognition problem. The AlexNet was trained on ImageNet dataset which took as input 224×224 pixels RGB images and categorized it into the respected class. This architecture consists of five convolutional layers from C1-C5 and three fully connected layers Fc6-Fc8 as shown in the top row of Figure 4.1.

However, this architecture contains 60 million parameters, learning a large number of parameters for small training dataset of the new task is problematic and time consuming. Therefore, a source architecture has been used as a feature extractor followed by an off-the-shelf hybrid SVM-KNN classifier for action recognition. The value of K in the nearest neighbor algorithm is selected through cross validation. The proposed work is innovative and presents an interesting combination of deep learning and hybrid classifier, which results in boosting the performance of the human recognition method. The experimentation

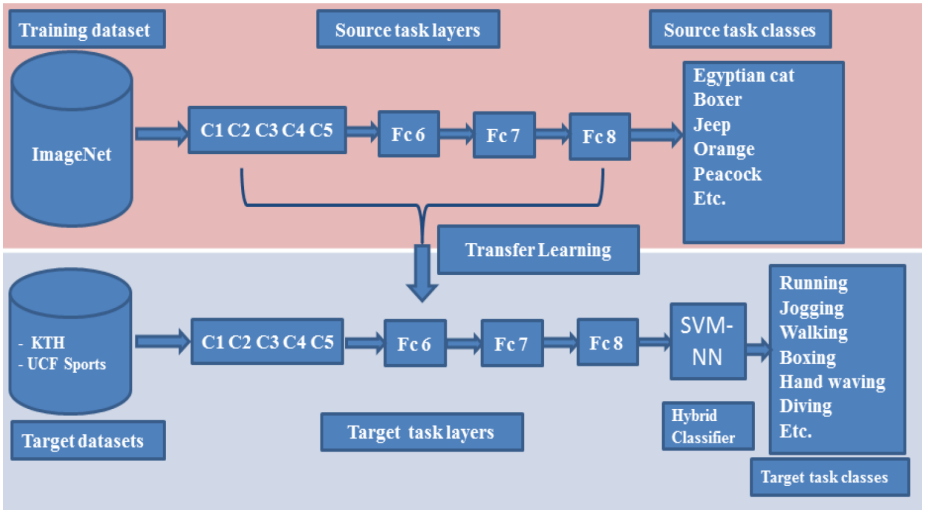


FIGURE 4.1: Overview of the proposed system, first row indicates the source architecture and second row shows the target architecture

results confirm the efficiency of the proposed work. Moreover, it is also demonstrated that the hybrid classifier has an advantage over a single classifier. The hybrid classification model based on SVM-KNN is presented in Figure 4.2.

4.4 Experimentations and Results

This section discusses the experimental setup, training process and experimental results for the proposed technique. The proposed technique is tested on two well-known action datasets i.e., KTH [15], and UCF Sports [17]. The description of these datasets and comparative analysis are presented in the subsequent sections.

4.4.1 Evaluation on the KTH dataset

The KTH [15] is well-known public dataset comprised of 6 actions, including waking, running, jogging, hand waving, boxing, and hand clapping. There were 25 actors involved in performing these actions in different setups including: outdoor, outdoor with variation in scale, outdoor with different clothes, and outdoor with illumination variations. The

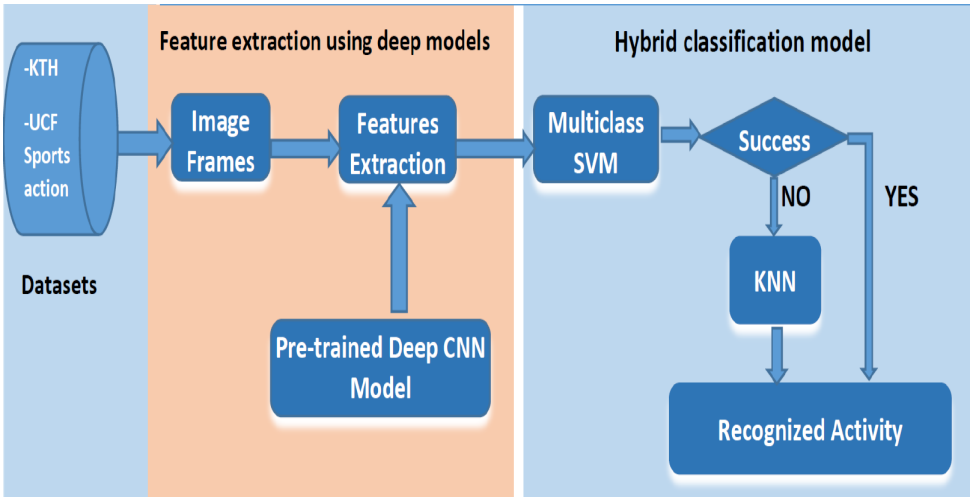


FIGURE 4.2: Feature extraction and hybrid classification model

sample frames for each action from four different scenarios are shown in Figure 4.3. This is a single view dataset with uniform background and recorded with fixed camera at the frame rate of 25fps.

During experimentation, the dataset is divided into two parts; one part is used for training, while other one is used for testing the correctness of the proposed method, same as [248]. The proposed method achieves 98.15% accuracy on KTH dataset, which is higher than the similar methods such as [73, 114, 164, 173, 248, 254, 255], as shown in Table 4.1. The confusion matrix indicating the accuracy of each action, and correspondence between the target classes along x-axis and output classes along y-axis, is shown in Figure 4.4.

4.4.2 Evaluation on UCF sports action dataset

The UCF dataset [17] encompasses 10 sports actions collected from videos broadcasted on television channels such as ESPN and BBC. These actions include: golf swing, diving, lifting, kicking, running, riding horse, swing-bench, skateboarding, swing-side, and

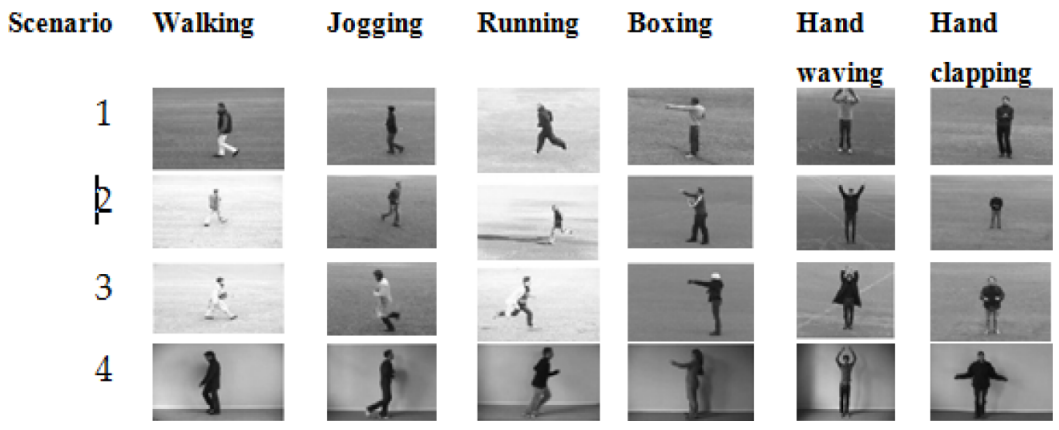


FIGURE 4.3: Sample frames for each action from four scenarios in KTH dataset

TABLE 4.1: Comparison of classification results on KTH dataset

Year	Method	Accuracy (%)
2017	**Proposed method (SVM-KNN)**	**98.15**
2017	**Proposed method (KNN)**	**94.83**
2017	**Proposed method (SVM)**	**89.91**
2016	Charalampous,and Gasteratos [248]	91.99
2016	Ahad et al. [254]	86.7
2016	Ding and Qu [255]	95.58
2013	Wang et al. [100]	94.2
2013	Ji et al. [173]	90.2
2013	Chaaraoui et al. [114]	89.86
2011	Le et al. [164]	93.9

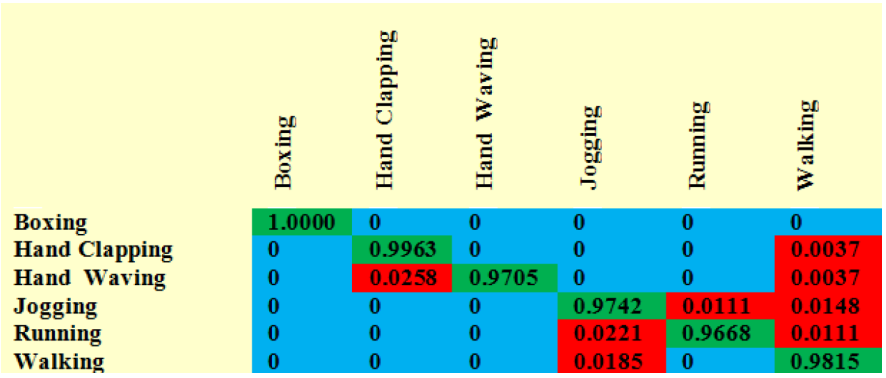


FIGURE 4.4: Confusion matrix of KTH dataset with 6 human actions

walking. These actions were recorded in a real sport environment exhibiting the variations in background, illumination conditions, and occlusions, which make it a challenging dataset. The sample frames for each action are shown in Figure 4.5.

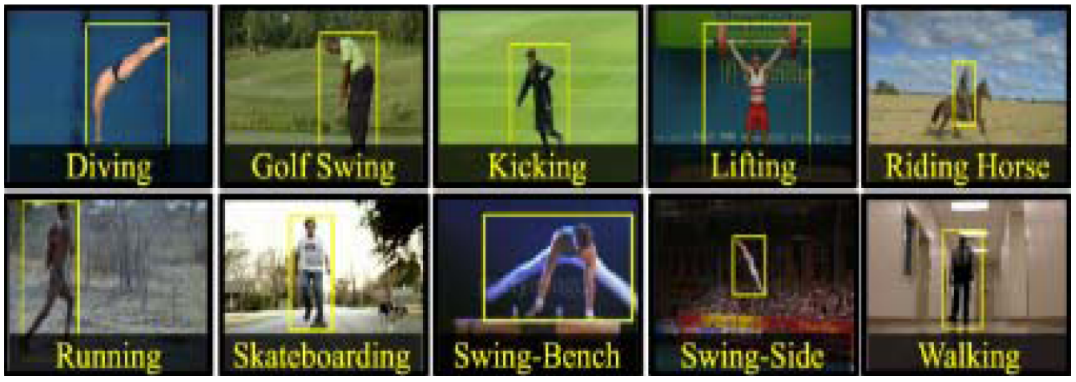


FIGURE 4.5: Sample frames for each action from UCF sports dataset

The proposed study uses a popular Leave-One-Out (LOO) cross validation scheme. Some other methods have also used Leave-One-Sequence-Out (LOSO), and Leave-One-Person-Out (LOPO) cross validation, which are quite similar to LOO validation [112]. In LOO cross validation, all video sequences are used for training except one, which is used for testing the performance of the classifier. This method is repeated for all available video sequences. Finally, the results of these sequences are summed up and the average result is considered as the final result. This validation scheme has been employed by many similar research methods, such as [100, 256], for assessing the performance of their methods. Since, the proposed method uses the same validation scheme, it provides a fair comparison with similar methods. The proposed transfer learning method achieved an accuracy of 91.47% on UCF sports dataset, which is higher than other similar methods as shown in Table 4.2. The detailed confusion matrix indicating the accuracy of each action, and correspondence between the target classes along x-axis and output classes along y-axis is shown in Figure 4.6.

TABLE 4.2: Comparison of classification results on UCF sports action dataset

Year	Method	Testing Scheme	Accuracy (%)
2017	**Proposed method (SVM-KNN)**	**LOO**	**91.47**
2017	**Proposed method (SVM)**	**LOO**	**89.60**
2017	**Proposed method (KNN)**	**LOO**	**82.75**
2016	Tian et al. [256]	LOO	90.0
2016	Charalampous and Gasteratos [248]	-	88.55
2015	Atmosukarto et al. [257]	LOO	82.6
2014	Yuan et al. [245]	LOO	87.33
2013	Wang et al. [100]	LOO	88.0
2011	Le et al. [164]	-	86.5
2011	Wang et al. [68]	LOO	88.2
2010	Kovashka et al. [258]	LOO	87.27
2009	Wang et al. [259]	LOO	85.6

	Diving	Golf swing	Kicking	Lifting	Riding horse	Running	Skateboarding	Swing-bench	Swing-side	Walking
Diving	1.00000	0	0	0	0	0	0	0	0	0
Golf swing	0	0.9928	0.0036	0	0	0.0018	0.0018	0	0	0
Kicking	0	0.0109	0.7464	0.0036	0.0145	0.0978	0.0399	0.0036	0.0072	0.0761
Lifting	0	0	0	1.0000	0	0	0	0	0	0
Riding horse	0	0	0	0	0.9094	0.0036	0	0	0.0833	0.0036
Running	0	0.0634	0.1178	0	0.0054	0.7482	0.0018	0.0127	0.0072	0.0435
Skateboarding	0	0	0.0326	0	0.0236	0.0272	0.9112	0	0.0054	0
Swing-bench	0	0	0	0	0	0	0	1.0000	0	0
Swing-side	0	0	0	0	0	0	0	0	1.0000	0
Walking	0	0.0091	0.0598	0.0036	0.0254	0	0.0562	0.0054	0.0018	0.8388

FIGURE 4.6: Confusion matrix of UCF sports action dataset

4.5 Conclusion

This chapter presented a human action recognition method based on transfer learning using a pre-trained deep CNN architecture and a hybrid SVM-KNN classifier. The source architecture is used as a feature extractor machine for the new task and hybrid SVM-KNN classifier is trained on the target datasets. It was experimentally confirmed that with the help of transfer learning we can successfully utilize the already learned knowledge for learning a new task with a limited training dataset. Transfer learning is very useful when the dataset is not sufficient for training the deep learning model from scratch. Moreover, training a deep learning model from scratch requires time and computational resources, which can be saved using transfer learning. In addition to this, it was confirmed that a hybrid classifier has an advantage over the single classifier in boosting the accuracy of the recognition system. Moreover, unlike handcrafted representation based methods, the proposed approach is simpler and directly works with RGB images, thus, eliminating the need of pre-processing and manual feature extraction. The effectiveness of the proposed method was confirmed on two well-known KTH, and UCF sports action datasets, and

achieved 98.15%, and 91.47% accuracy, respectively. The comparative analysis confirms that the proposed methods outperforms similar state-of-the-art methods for human action recognition using transfer learning. In future, the authors like to extend this method for more complex datasets such as IXMAS, UCF-50, UCF-101, and HMDB-51.

Chapter 5

Human Action Recognition using Deep Belief Networks

IN THIS CHAPTER

An innovative method for human action recognition using unsupervised Deep Belief Networks (DBNs) is presented. The introduction of the proposed method is presented in Section 5.1. While background of the DBNs model, and related work, are ascertained in Section 5.2. The proposed methodology and experimentations results are elaborated in Section 5.3 and 5.4 respectively. Finally, the chapter is concluded in Section 5.5.

5.1 Introduction

Over the past few years, vision based human activity recognition (HAR) has caught the attention of the research community due to its important applications in real world scenarios. A lot of methods have been proposed for human action recognition using a deep learning approach. Nevertheless, most of these methods are based on supervised

learning and little attention has been paid towards unsupervised deep learning models. However, as matter of fact, unsupervised learning is far more important than its supervised counterpart [38] since we discover the world by observing it rather being told the name of every object. Human and animal learning is mostly unsupervised.

Video contents on the internet are increasing rapidly. Every minute, 100 hours of videos are uploaded onto YouTube alone and according to the reports by 2017, 74% of the all internet traffic will consist of video content. Most of these video contents are not annotated, hence we cannot make sense out of them with supervised learning, until and unless these are annotated. However, the manual annotation process is time consuming and expensive. This rapid growth of video content demands for efficient and effective ways of understanding the video content. On the other hand, search engines use the text meta-data such as title, description, and tags provided by the users for searching and processing videos on the internet. These meta-data are incomplete, sparse, and in some cases inconsistent with the content of the videos. As a consequence, automatic video understanding techniques such as classification, retrieval, and analysis are inevitable. The great variety of videos uploaded on the internet or recorded by surveillance cameras represent realistic scenes and activities. These videos differ significantly from the videos which contain simple activities recorded under controlled settings. Most of the existing techniques for HAR target these simple activities such as in Weizmann [10] and KTH [15] datasets. However, datasets such as UCF sports [260], UCF11 [89], and UCF50 [201] are collected from different TV channels and YouTube videos and represent realistic and unconstrained activities. A good number of research articles have been published on HAR

using handcrafted and deep learning based techniques, but recognizing actions from unconstrained videos is still a challenging task and requires more attention.

There are two major approaches to address this problem, these include the traditional handcrafted feature-based, and the deep learning-based approach. The handcrafted approach suffers from several limitations such as, its inability to automatically learn features from the data, and shallow representation of data in classifiers. On the other hand, a deep learning-based approach employs the concept of end-to-end learning by using the trainable feature extractor followed by a trainable classifier. Deep learning has emerged as a popular approach within machine learning aimed at automatically extracting multiple layers of features from the raw data. Deep learning algorithms develop multi-layer representation of different patterns in the data, where each successive layer is responsible for learning increasingly complex features. Higher level features are learned from the input of the lower layers, thus representation increased abstraction level at each consecutive layer. This automatic learning ability of deep learning models eliminates the need of handcrafted feature detectors and descriptors. In addition to this, these models have shown superior performance over traditional handcrafted feature-based techniques in many visual categorization tasks [19].

The deep learning model such as Convolutional Neural Networks (CNNs) [230] and Deep Belief Networks (DBNs) [261], deep recurrent neural networks [262, 263], and deep Boltzmann machines [264] have been successfully employed for many visual categorization tasks. Among these models, DBN is an unsupervised probabilistic graphical model capable of learning from the input data without any prior knowledge. This model

can also be trained in a semi-supervised or unsupervised fashion which is quite helpful when we have few samples of labeled data or we are dealing with unlabeled data. As discussed earlier most of the available videos are in unlabeled form, and annotating the data manually is time consuming and expensive.

In this chapter, an unsupervised deep learning based method for human action recognition in unconstrained videos is proposed. This method employs Deep Belief Networks with restricted Boltzmann machine to construct a deep neural network model. This method takes raw images data as input and automatically learns the feature representation suitable for classification of human actions.

5.2 Background and Related Work

Deep learning models have been under consideration since late 60s [159], however, the importance of these models was overshadowed by the success of shallow networks such as Support Vector Machines (SVMs) [265]. In recent years, the achievement of these models was confirmed with the introduction of an efficient Deep Belief Network (DBN) model in [161]. This is a multi-layer generative model where each higher layer represents more abstract information. In the first phase, the model is pre-trained using multiple layers of Restricted Boltzmann Machines (RBMs), then in the second phase, parameters are fine-tuned in a supervised manner using backpropagation [266, 267]. With efficient and successful training of DBN, deep learning models have become popular, and applied to many applications such as object recognition [268, 269], image classification [270], face image analysis [271], visual task categorization [272], and speech recognition [263].

5.2.1 Restricted Boltzmann Machine

The Restricted Boltzmann Machine [163] is a probabilistic graphical model originated from Deep Boltzmann Machine (DBM) [273]. This model is used by DBN for learning each layer of representation in an unsupervised manner. The original Boltzmann machine consist of a visible or input layer followed by many hidden layers and imposes no constraints for connecting any type of nodes. However, to make the training process simpler and easier, certain restrictions are imposed on RBM such as it has no links between nodes of the same type, making it a bi-partite graph. Hence, this model consists of a visible layer followed by hidden layers, where symmetric connections exist between the layers but not within the layers as shown in Figure 5.1. As a consequence, RBM model is much quicker to train than original the DBM due to conditional independence of the hidden layers. Therefore, RBM is preferred over the DBM model for training the deep belief networks.

The RBM and its variants have been applied to many applications such as documents retrieval [274], visual categorization [162, 275–277], motion data modeling [278], human activity recognition [279, 280], speech and audio data analysis [281, 282]. In visual categorization tasks, the visible or input layer corresponds to the input (e.g., image pixels), and hidden layers correspond to the feature detectors as a collections of neurons in the visual corridor. The RBM either uses the binary visible and hidden nodes for binary input or Gaussian visible nodes for continuous valued data. The subsequent sections discuss how the RBM model is trained with different types of visible nodes.

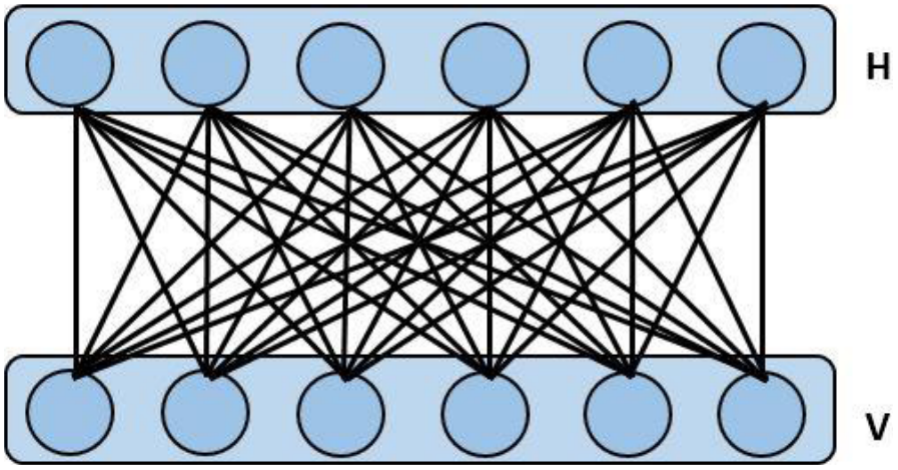


FIGURE 5.1: Restricted Boltzmann Machine model with visible layer (V) and hidden layer (H). The connections exist between the layers but not within the layers. The visible layer represents the input given to the model and hidden layer is used to learn the features from the input data.

5.2.1.1 RBM with binary visible nodes

In the RBM model, the probability assigned to the input layer and hidden layers configuration (v, h) can be calculated using the energy function [283]. In the case of binary visible and hidden nodes, the energy function is formulated as:

$$E(v,h) = -a^T v - b^T h - v^T W h \tag{5.1}$$

where v $\in \{0,1\}^n$ and h $\in \{0,1\}^m$ representing the states of the input and hidden nodes, respectively. The W $\in \mathbb{R}^{nxm}$ is a weight matrix that indicates the symmetric connection between the input and hidden nodes, while a $\in \mathbb{R}^n$ and b $\in \mathbb{R}^m$, indicate the biases for input and hidden nodes, respectively. The probability of joint configuration $\{$v, h$\}$ can be calculated as:

$$p(v,h) = \frac{e^{-E(v,h)}}{\sum_{\eta,\mu} e^{-E(\eta,\mu)}} \tag{5.2}$$

and probability assigned by the RBM model with configuration v, h, can be obtained by marginalizing h, as:

$$p(v) = \frac{\sum_h e^{-E(v,h)}}{\sum_{\eta,\mu} e^{-E(\eta,\mu)}} \tag{5.3}$$

In unsupervised RBM learning, a log probability is increased for the training data, which reduces the energy of the training data. During the training phase, this is achieved by adjusting the weights and biases values. In this direction, the log probability of the training data can be maximized using stochastic steepest ascent rule as:

$$\Delta w_{ij} = \varepsilon \left(< w_i h_j >_{data} - < v_i h_j >_{model} \right) \tag{5.4}$$

where ε is a learning rate and $< . >_{\Phi}$ is an expectation under the distribution. The probability of RBM with binary visible and hidden nodes h_j and randomly chosen training sample v can be calculated as:

$$p\left(h_j = 1 | \, v\right) = \frac{1}{1+\exp(-b_j - \sum_i v_i w_{ij})} \tag{5.5}$$

This is a logic sigmoid function applied to the total input of the hidden nodes. The example of an unbiased sample is given by $v_i h_j$. The probability of the visible node v_i being 1, given the state h of hidden nodes is

$$p\left(v_i = 1 | \, h\right) = \frac{1}{1+\exp(-a_i - \sum_j h_j w_{ij})} \tag{5.6}$$

However, calculating the unbiased sample of $(v_i h_j)_{model}$ would require a long sequence

of alternating Gibbs sampling between the visible and hidden layers, therefore mostly the approximation rule is applied.

5.2.1.2 Contrastive Divergence Learning

Contrastive Divergence Learning (CD) [284] is an efficient approximate training method for RBM. Although CD only approximates the log probability of the training data, in practice CD has proven to be quite helpful in producing good model. Stacked RBMs trained using CD are powerful tools for pre-training the DBN as a generative model. This model has been successfully applied to visual categorization, and other tasks that involve complex data. The single step version of CD algorithm works as follows:

1. First, the visible states are initialized with training samples.
2. Binary hidden states are sampled in parallel, according to Equation 5.5.
3. Visible states are sampled according to the Equation 5.6. This is known as the reconstruction step.
4. Finally, the weights are updated using Equation 5.7.

$$\Delta w_{ij} = \varepsilon \left(< v_i h_j >_{data} - < v_i h_j >_{reconst} \right) \tag{5.7}$$

However, in order to improve the performance of the model, the sampling may be repeated multiple times in each step. This is as a general form of the CD algorithm CDn, where n shows the number of iterations.

5.2.1.3 RBM with Gaussian visible nodes

The RBM model with Gaussian visible nodes is more difficult to train than with binary visible nodes. However, with continuous valued data, Gaussian nodes provide better

performance. In this case, the energy function is calculated as:

$$E\ (v,h)\ = \sum_i \frac{(v_i - a_i)^2}{2\sigma^2 i} - \sum_j b_j h_j - \sum_{i,j} \frac{vi}{\sigma i} h_j w_{ij} \tag{5.8}$$

and probability of hidden node activation is calculated as:

$$p\left(h_j = 1 \mid v\right) = \frac{1}{1 + \exp(-b_j - \sum_i \left(\frac{vi}{\sigma i}\right) w_{ij})} \tag{5.9}$$

and the value of Gaussian visible nodes during reconstruction step is calculated as:

$$< vi >_{reconst} = a_i + \sigma_i \sum_j h_j w_{ij} \tag{5.10}$$

5.2.2 Deep Belief Networks

Deep Belief Networks (DBNs) are graphical models with multi-layer architecture and can work as a generative or unsupervised model. The multiple layers of DBN enables automatic feature extraction in a non-linear fashion even from complex data distributions. This model can be trained effectively using [161]. In the first phase, the model is pre-trained in a greedy layer-wise fashion using unsupervised RBMs for learning network weights and other parameters. In this layer-wise training, once the hidden layer is trained, the weights for this layer become fixed, and activation on this hidden layer provides input to the next RBM and so on. In the second phase, the model can be fine-tuned in a generative or discriminative manner to solve a specific task. The fine-tuning process can be conducted using backpropagation or a variant of the wake-sleep algorithm [161, 285].

An example diagram of DBN with binary visible nodes is shown in Figure 5.2.

5.2.2.1 Generative form of DBNs

The DBN generative model consists of multiple consecutive layers, where lower layers have directed connections, and top two layers have undirected connections and form an associative memory. Moreover, adding a further hidden layer to a generative DBN model may improve the performance of the model if certain pre-conditions are met [161]. The fine-tuning of a generative model can be conducted using the wake-sleep algorithm [286].

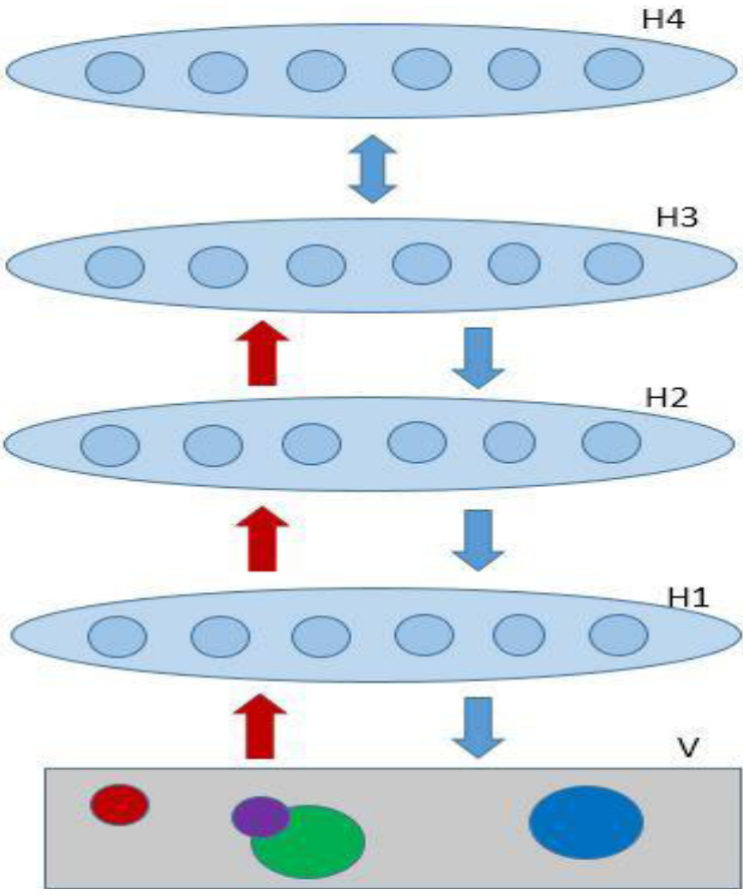


FIGURE 5.2: Diagram of DBN with visible layer V and four hidden layers (H1-H4). The blue arrows indicate the generative model, while red arrows indicate the direction of recognition.

After the pre-training phase, a generative model can be trained in a completely unsupervised manner, or labels can be used for training the model. The model proposed in [285] learns to generate new data along with appropriate class labels. In this connection, labels are introduced at penultimate layer of DBN, and the top RBM layer is used for learning using these labels as shown in Figure 5.3. This trained model can be used for a classification task, in this case, after an input the bottom-up activations are calculated. However, bottom-up activations on the penultimate layer are kept fixed and Gibbs sampling is conducted between the top two layers and the probability of each class label is calculated.

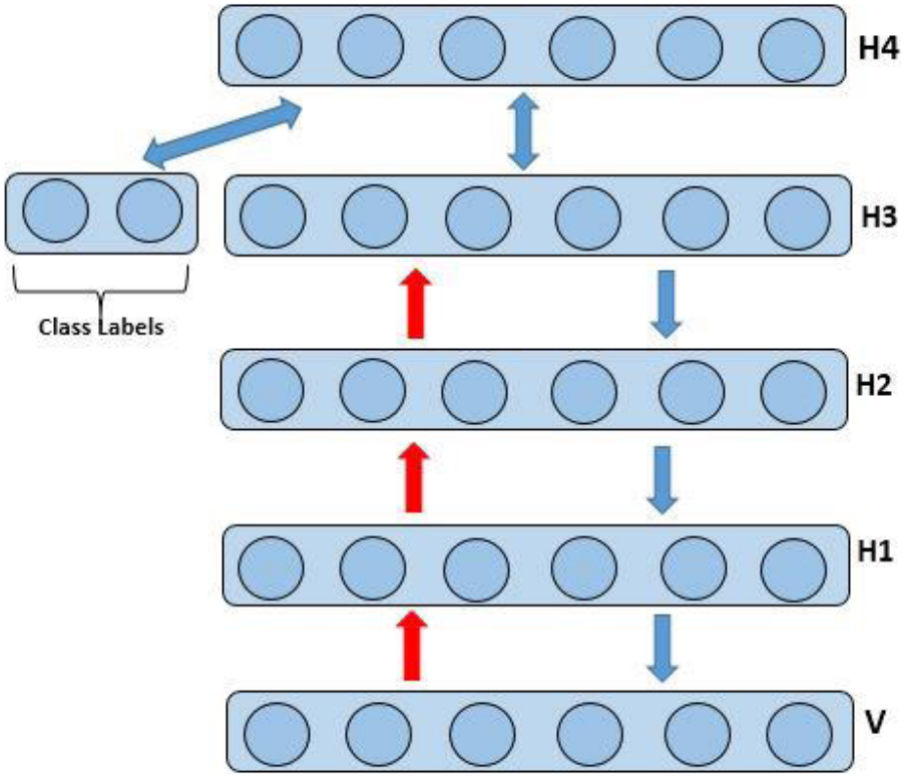


FIGURE 5.3: The DBN generative model with four hidden layers (H1-H4), used for generating images with their class labels.

5.2.2.2 Discriminative form of DBNs

First, the DBN model is trained in a layer-wise fashion as a generative model, then an output node is added with class labels, and model is tuned using backpropagation to solve the classification task. The backpropagation is performed on a training set to minimize the classification error, which can be measured using cross-entropy error:

$$[H\,(p,\ \hat{p}) = \ -\ \sum_{i} pi\ln\hat{p}i \tag{5.11}$$

where p indicates the training label of each category, i and $\hat{p}_i$ denotes the probability of the training example belonging to the category i as predicted by the model. An example diagram of the discriminative DBNs is shown in Figure 5.4. The discriminative DBNs model provide better performance than the generative models on classification tasks [285].

5.2.3 Structure Learning for Deep Networks

Structural learning can improve the performance and learning capability of the model by automatically identifying the set of best parameters during the training process. These parameters describe the architecture of the model such as the number of layers, number of nodes in each layer, and learning rate. In this regard, a number of approaches have been proposed in the literature for structure learning of deep belief networks and other deep learning models. However, due to scalability issues, the use of structural learning for complex real world problems is limited. The most promising methods of structure learning are based on Bayesian non-parametric networks. These methods enforce prior

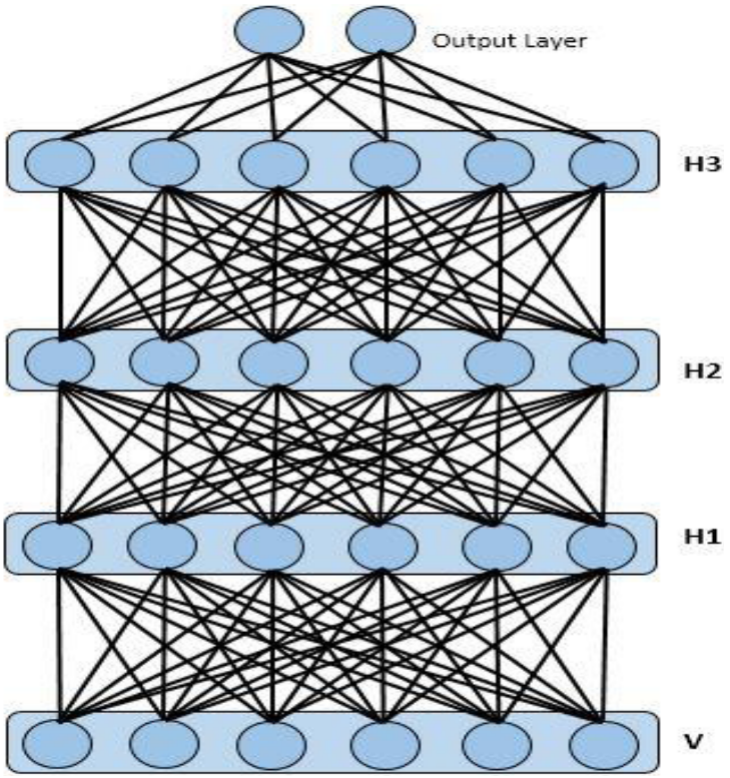


FIGURE 5.4: The discriminative DBN model for classification with a visible layer v, 3 hidden layers (H1-H3), and an output layer.

probability distribution over the structural components of the model, and use a probabilistic inference method to infer the structure and parameters of the model. A popular method for structure learning using Bayesian non-parametric methods was proposed in [287]. In the case of deep belief networks, heuristic methods [288] and evolutionary methods [289] have been proposed in the literature for structure learning. For a specific class of neural networks, a method proposed in [290] can adapt the networks structure in an online manner.

5.3 Proposed Methodology

In recent years, various methods have been proposed for human action recognition based on a handcrafted and deep learning based approach. However, deep learning based methods are gaining popularity due to their capability of automatic feature extraction and representation. Recently, there has been increased interest in deep architectures such as DBNs [161]. The goal of these models is to learn more abstract representations from the input data in a layer-wise fashion using a pre-training unsupervised learning phase. After pre-training, this model can be trained in a completely unsupervised manner or labels can be introduced at penultimate layer of the model as discussed in section 5.2.2.1. The proposed framework for human action recognition is based on DBN with multiple layers of Gaussian-Bernoulli RBM for modelling the complex human activities in unconstrained videos.

The major focus of the proposed method is to learn the effective feature representation suitable for action recognition. In this direction, we used multiple layers of RBMs using greedy layer-wise pre-retraining by contrast divergence method. The greedy layer-wise pre-training method is quite helpful in initializing weights of the DBN, which increases the learning capability of the model and avoids over-fitting. This pre-training step generates many layers of feature detectors and descriptors suitable for feature representation. After the pre-training phase, we used backpropagation fine tuning which slightly modified the feature detectors and descriptors according to the class labels and made them suitable for discrimination tasks. The major advantage of the pre-training phase is that the labels

information is not used for the features representation from scratch, such as in supervised deep learning models.

However, it has been observed that training and fine-tuning of the DBN, with multiple RBMs layer, is a time consuming task. Therefore, we resized the extracted images from the video frames into 100x100 size, which made 10000 dimensions, as the actual input to the model at the visible or input layer. In addition to this we used 3 hidden layers and an output layer, which corresponds to the 10 output classes of the model as shown in Figure 5.5. The parameter configuration of the model is shown in Table 5.1. We performed experiments using MATLAB R2017a implementation on graphics works station with Intel Xeon Processor E5-2630 v3 (8C, 2.4GHz, Turbo, HT, 20M, 85W), 64GB RAM, NVIDIA Quadro, K2200 4GB GPU.

5.4 Experimentation and Results

In this section the details of the experimental setup, training process and experimental results of the proposed method are discussed. The performance of the proposed method

TABLE 5.1: Training parameters for proposed model

S#	Parameter	Value
1	No. of total layers	5
2	No. of hidden layers	3
3	Unsupervised learning rate	0.001
4	Supervised learning rate	0.1
5	No. of unsupervised epochs	100
6	No. of supervised epochs	300
7	Initial momentum	0.5
8	Final momentum	0.9
9	Batch size	100
10	Training accuracy	100%
11	Testing accuracy	88.86%

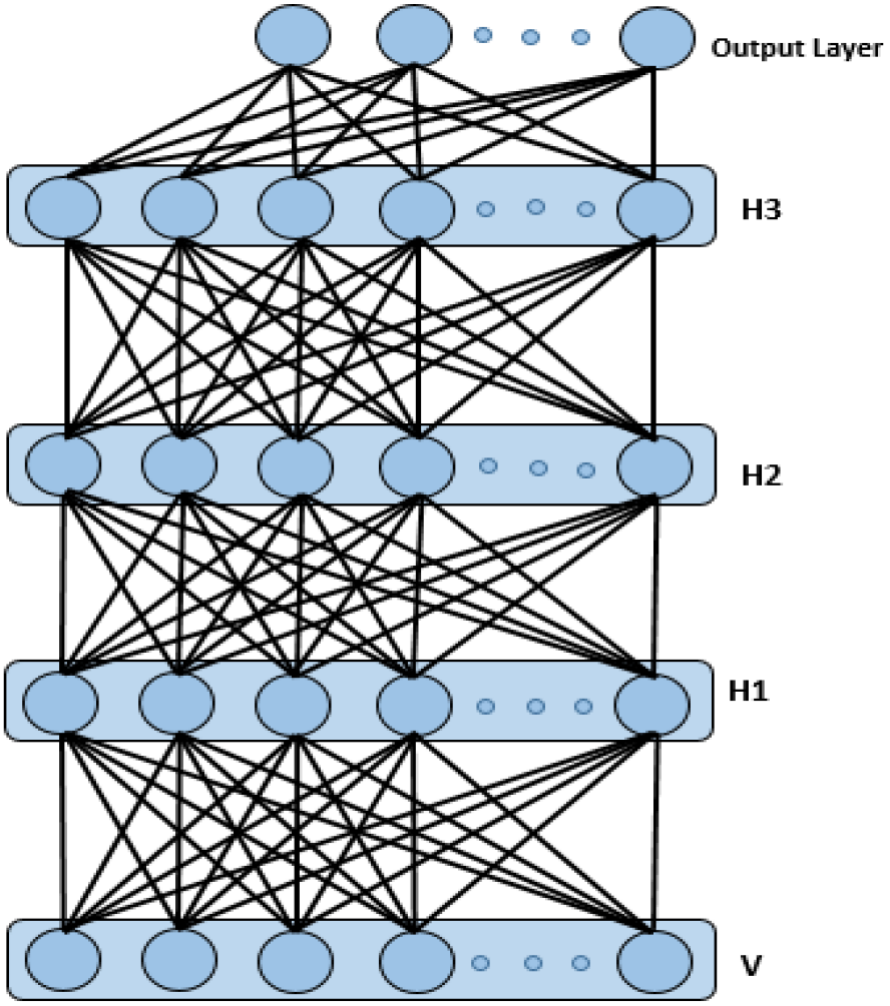


FIGURE 5.5: The proposed DBN model having visible layer v with 100x100 dimension input, 3 hidden layers (H1-H3), and an output layer with 10 nodes representing 10 action classes.

has been assessed over the UCF Sports dataset [16]. This dataset is comprised of 10 sports actions collected from variety of sports, broadcasted on television channels such as BBC and ESPN. Actions include: diving, golf swing, kicking, lifting, riding horse, running, skateboarding, swing-bench, swing-side, and walking. These actions were recorded in a real sport environment exhibiting the variations in background, illumination conditions, and occlusions, which make it a challenging dataset. The sample frames for each action are shown in Figure 5.6.

FIGURE 5.6: Sample frames for each action from UCF sports dataset

TABLE 5.2: Comparison of classification results on UCF sports action dataset

Year	Method	Validation Scheme	Accuracy (%)
-	**Proposed method**	**LOO**	**88.86**
2015	Atmosukarto et al. [257]	LOO	82.6
2014	Yuan et al. [245]	LOO	87.33
2013	Wang et al. [100]	LOO	88.0
2013	Tian et al. [291]	-	75.2
2011	Wang et al. [68] (Trajectory)	LOO	75.2
2011	Le et al. [164]	-	86.5
2010	Kovashka and Grauman [258]	-	87.27

The proposed method uses a popular Leave-One-Out (LOO) cross validation scheme, as recommended in pioneering work on this dataset [260]. In this scheme, one action sequence is used for testing, while rest of the action sequences are used for training the classifier [21]. This process is repeated for all action sequences and resultant accuracies are averaged. This validation scheme has been adopted in [68, 100, 245, 257, 291] for the evaluation of their proposed techniques. This shows fair comparison of our model with these state-of-the-art methods. The proposed unsupervised deep belief networks based method achieved an average accuracy of 88.86% on this dataset, which is higher than similar state-of-the-art methods as shown in Table 5.2.

5.5 Conclusion

This chapter presented an unsupervised deep belief networks based method for recognition of human activities in unconstrained videos. The proposed method automatically extracts suitable feature representation without any prior knowledge using an unsupervised deep learning model. In this work, the contribution is twofold. Firstly, the results indicate that the proposed method achieved higher accuracy than state-of-the-art unsupervised deep learning based methods. This demonstrates the potential of less explored unsupervised DBNs architecture for modelling complex human activities. Secondly, although the model has been designed by keeping human activities in mind, and its effectiveness is confirmed by high recognition results on a challenging UCF sports dataset,this model can also be adapted for other visual recognition tasks as well. As future work, this method can be extended for unstructured human activities such as crowed behavior analysis using unsupervised DBN model, since the major portion of video content available on the internet are unlabeled and unstructured videos.

Chapter 6

Conclusion and Future Research Directions

IN THIS CHAPTER

The conclusive summary, research objectives, and extensions related to the proposed techniques are presented. In this regard, research objectives and their realization is presented in Section 6.1. The possible extensions and enhancements of the proposed techniques in the short-term and long-term perspectives are presented in Section 6.2

6.1 Research objectives and their realization

In this section, the research objectives (O1 to O7) are listed as pre-specified in chapter 1 and their realization is demonstrated briefly.

O1. Comprehensive review of the state-of-the-art techniques based on handcrafted and deep learning approaches, in order to decide which one works the best.

O2. Understanding the limitations of the state-of-the-art techniques, and identify the gaps for contributions.

These two objectives have been realized by conducting a comprehensive review of the state-of-the-art techniques. In this regard, most popular and prominent techniques were reviewed, compared, evaluated, and analyzed on different well known public datasets. The key findings and analysis regrading handcrafted and deep learning based techniques are presented in chapter 2 and briefly discussed as follows:

Handcrafted-based techniques

In this category, Space-Time-based techniques have been popular and have achieved promising results on simple and complex datasets. Among these techniques, trajectory and space-time feature based methods have shown overall superior performance which are also reliable under noise and different illumination conditions. However, these techniques also have some limitations. Firstly, correct localization of subject and movement tracking is still a challenging task in these methods. Secondly, these methods are computationally expensive, and may not be suitable for real-time applications. Therefore, by considering the nature of the problem in hand, one may resort to other techniques including appearance, and region-based methods.

Deep Learning-based techniques

Deep learning has emerged as a highly popular direction within machine learning, which has outperformed the traditional approaches in many applications of computer vision. The highly advantageous property of deep learning algorithms is their ability to

learn features from the raw data, which eliminates the need of handcrafted feature detectors and descriptors. However, this approach also has some limitations. Deep learning models require a huge amount of data for training the algorithm. However, most of the well-known public HAR datasets such as KTH [15], IXMAS [141], HMDB-51 [72], and UCF Sports [16, 17] are small in size, and therefore not suitable for training deep learning models from scratch.

Based on the above findings, major gaps were identified, and novel methods for human action recognition have been developed based on handcrafted and deep learning-based approaches.

O3. Development of a novel method for view-invariant human action, which is considered as one of the major challenges for recognition of human actions in different application domains.

This objective has been realized by developing a novel method for human action recognition from multiple views presented in chapter 3, and briefly discussed here. This method employs region-based features extracted from the human silhouette obtained from videos. The human silhouette contains much less information than the original image frame. However, it was experimentally proved that this information is sufficient to recognize the action with high accuracy, as this method achieved high accuracy compared to state-of-the-art handcrafted based methods, including trajectory based methods. In this work, two goals are achieved. Firstly, the proposed method achieves higher recognition accuracy than state-of-the-art methods on well-known multiview IXMAS (INRIA Xmas Motion Acquisition Sequences) dataset. Secondly, the average testing speed of the proposed

method is 34 frames per second, which is much higher than existing methods, and shows its suitability for real-time applications. The comparison with state-of-the-art methods in terms of accuracy is shown in Table 6.1, and in terms of speed of execution is shown in Table 6.2.

O4. Development of an innovative method for human action recognition using supervised deep learning or transfer learning model.

This objective has been realized by developing an innovative method for HAR based on transfer learning presented in chapter 4, and briefly discussed here. As mentioned earlier, deep learning models cannot be trained effectively with a small dataset. To overcome this limitation, and benefit from the best properties of deep learning models, a transfer learning approach was adopted. This method uses a pre-trained Convolutional Neural

TABLE 6.1: Comparison with state-of-the-art methods

Year	Method	Accuracy (%)
2016	**Proposed method**	**89.75**
2016	Chun et al. [118]	83.03
2013	Chaaraoui et al. [114]	85.9
2011	Wu et al. [63]	89.4
2011	Junejo et al. [206]	74
2010	Weinland et al.[211]	83.4
2009	Reddy et al. [226]	72.6
2008	Liu and Shah [213]	82.8
2008	Cherla et al. [229]	80.1
2008	Vitaladevuni et al. [228]	87.0
2007	Lv and Nevatia [227]	80.6

TABLE 6.2: Comparison with state-of-art methods for speed of execution

Method	Average FPS
Proposed method	**34**
Chaaraoui et al. [114]	26
Cherla et al. [229]	20
Lv and Nevatia [227]	5.1

Networks (CNNs) model as a source architecture for extracting the features from the target dataset, followed by a hybrid Support Vector Machines and K-Nearest Neighbor (SVM-KNN) classifier for action recognition. The proposed method was evaluated on two well-known action datasets, i.e., UCF sports and KTH. The comparative analysis confirmed the superior performance of the proposed method as shown in Table 6.3, and 6.4.

TABLE 6.3: Comparison with state-of-the-art methods on KTH dataset

Year	Method	Accuracy (%)
2017	**Proposed method (SVM-KNN)**	**98.15**
2017	**Proposed method (KNN)**	**94.83**
2017	**Proposed method (SVM)**	**89.91**
2016	Charalampous,and Gasteratos [248]	91.99
2016	Ahad et al. [254]	86.7
2016	Ding and Qu [255]	95.58
2013	Wang et al. [100]	94.2
2013	Ji et al. [173]	90.2
2013	Chaaraoui et al. [114]	89.86
2011	Le et al. [164]	93.9

TABLE 6.4: Comparison with state-of-art method on UCF sports action dataset

Year	Method	Testing Scheme	Accuracy (%)
2017	**Proposed method (SVM-KNN)**	**LOO**	**91.47**
2017	**Proposed method (SVM)**	**LOO**	**89.60**
2017	**Proposed method (KNN)**	**LOO**	**82.75**
2016	Tian et al. [256]	LOO	90.0
2016	Charalampous and Gasteratos [248]	-	88.55
2015	Atmosukarto et al. [257]	LOO	82.6
2014	Yuan et al. [245]	LOO	87.33
2013	Wang et al. [100]	LOO	88.0
2011	Le et al. [164]	-	86.5
2011	Wang et al. [68]	LOO	88.2
2010	Kovashka et al. [258]	LOO	87.27
2009	Wang et al. [259]	LOO	85.6

O5. Development of an innovative method for human action recognition using unsupervised deep learning model.

This objective has been realized by developing an innovative method for human action recognition based on the unsupervised deep learning method presented in chapter 5, and briefly discussed here. Deep learning based methods have gained popularity due to their high recognition results in different computer vision tasks. However, this success is mainly limited to the domains where a huge amount of annotated datasets are available. In other words, these success stories are mainly associated with supervised deep learning models. Due to these reasons, little attention has been paid towards unsupervised deep learning models. But, as matter of fact, unsupervised learning is far more important than its supervised counterpart [38], since we discover the world by observing it rather being told the name of every object. Human and animal learning is mostly unsupervised. Furthermore, the huge amount video content available on the Internet are unlabeled. Hence, we cannot recognize these contents with supervised learning techniques. The ultimate solution lies in developing and adopting unsupervised deep learning models.

Keeping in mind the importance of unsupervised deep learning, this thesis proposed an unsupervised deep belief networks based method for recognition of human activities in unconstrained videos. The proposed method automatically extracts suitable feature representation without any prior knowledge using an unsupervised deep learning model. In this work, our contribution is twofold. Firstly, the proposed method achieved higher accuracy than similar state-of-the-art methods, which confirmed the potential of less explored unsupervised DBNs architecture for modelling complex human activities as shown in Table

6.5. Secondly, the model uses the automatic structure learning method for learning important parameters. Although, the model has been designed by keeping human activities in mind, this model can also be adapted for other visual recognition tasks as well.

O6. Comparison between the supervised and unsupervised deep learning models on a same dataset.

This objective has been realized by evaluating the proposed supervised (transfer learning) and unsupervised deep learning-based methods on same dataset. It has been learned that, the proposed method using transfer learning produced better results 91.47%, as compared to our unsupervised deep belief network based method, which produced 88.6% accuracy. However, in cases where labeled data are rare or unavailable, then the unsupervised deep learning based methods prevails the supervised deep learning or transfer learning based methods. The comparison of the proposed methods with state-of-the-art methods is presented in Table 6.5 and Table 6.4.

O7. Production of better results than the existing ones in terms of accuracy and efficiency on standard benchmark datasets.

In fact, this was an ultimate objective of this research work. It has been realized by

TABLE 6.5: Comparison of classification results on UCF sports action dataset

Year	Method	Validation Scheme	Accuracy (%)
-	**Proposed method**	**LOO**	**88.86**
2015	Atmosukarto et al. [257]	LOO	82.6
2014	Yuan et al. [245]	LOO	87.33
2013	Wang et al. [100]	LOO	88.0
2013	Tian et al. [291]	-	75.2
2011	Wang et al. [68] (Trajectory)	LOO	75.2
2011	Le et al. [164]	-	86.5
2010	Kovashka and Grauman [258]	-	87.27

proposing a novel methods for human action recognition, which produced superior results than state-of-the-art methods as shown in Tables 6.1, 6.2, 6.3, 6.4, and 6.5.

6.2 Future Research Directions

6.2.1 Short-term perspective

This section presents the short-term plan for possible extensions of the proposed methods.

Extension of the handcrafted based method for cross-view action recognition

Human action recognition from multiple views is proposed in chapter 3. This method showed high accuracy and efficiency on a challenging dataset. In future, authors aim to extend this method for cross-view action recognition and evaluate it on recently developed multi-view and multi-modal datasets, such as [292].

Extension of transfer learning based HAR method

This innovative method for human action recognition is based on transfer learning. In this method, the pre-trained deep CNN architecture was used as a feature extraction machine and a hybrid SVM-KNN classifier was employed for classification. This is considered as a partial transfer learning. In future, authors would like to extend this method for more challenging datasets, such as [188], using deep/full transfer learning.

Evaluation of proposed methods on cross datasets

The proposed techniques were trained and tested on the same datasets. In future, authors would like to test proposed techniques on cross datasets, and see whether the performance decreases or remains same.

6.2.2 Long-term perspective

This section presents the long-term plan for improvement and adaptation of the proposed methods.

Extension of unsupervised deep learning based method

This method is based on unsupervised deep belief networks for human action recognition in unconstrained videos. The proposed method automatically extracts suitable feature representation, without any prior knowledge, using an unsupervised deep learning model. The proposed method achieved higher accuracy than similar state-of-the-art deep learning based methods. For future work, this method can be extended for unstructured human activities such as crowed behavior analysis using an unsupervised DBN model.

Extension of proposed deep learning based methods for real time action recognition

Most of the existing deep learning based methods are not suitable for real time action recognition due to their computational complexities. It is inevitable to make the necessary changes in existing methods to meet real time requirements. In future, authors plan to make the required changes in proposed deep learning based algorithms to make them suitable for real time action recognition.

Extension of proposed deep learning based methods for detecting suspicious events in videos

Automatic detection of suspicious events is very important for video surveillance. Most of the existing techniques for suspicious events detection are based on handcrafted appearance and motion features. However, these methods are not optimal for analysis of

complex video scenes. Authors would like to extend the proposed deep learning based methods for detection of suspicious events in real-time during video surveillance.

Bibliography

[1] A. Karpathy, F. Li, and J. Johnson, "Cs231n convolutional neural network for visual recognition,"," *Online Course*, 2016.

[2] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, pp. 568–576.

[3] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys (CSUR)*, vol. 43, no. 3, p. 16, 2011.

[4] L. Kratz and K. Nishino, "Tracking pedestrians using local spatio-temporal motion patterns in extremely crowded scenes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 5, pp. 987–1002, 2012.

[5] T. Xiang and S. Gong, "Video behavior profiling for anomaly detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 5, pp. 893–908, 2008.

[6] M. Mubashir, L. Shao, and L. Seed, "A survey on fall detection: Principles and approaches," *Neurocomputing*, vol. 100, pp. 144–152, 2013.

[7] H.-K. Lee and J.-H. Kim, "An hmm-based threshold model approach for gesture recognition," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 21, no. 10, pp. 961–973, 1999.

[8] L. Zhang, M. Jiang, D. Farid, and M. A. Hossain, "Intelligent facial emotion recognition and semantic-based topic detection for a humanoid robot," *Expert Systems with Applications*, vol. 40, no. 13, pp. 5160–5168, 2013.

[9] L. Xia, I. Gori, J. K. Aggarwal, and M. S. Ryoo, "Robot-centric activity recognition from first-person rgb-d videos," in *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pp. 357–364, IEEE, 2015.

[10] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[11] Y. Luo, T.-D. Wu, and J.-N. Hwang, "Object-based analysis and interpretation of human motion in sports video sequences by dynamic bayesian networks," *Computer Vision and Image Understanding*, vol. 92, no. 2, pp. 196–216, 2003.

[12] R. M. Vallim, J. A. Andrade Filho, R. F. De Mello, and A. C. De Carvalho, "Online behavior change detection in computer games," *Expert Systems with Applications*, vol. 40, no. 16, pp. 6258–6265, 2013.

[13] S. G. Klauer, F. Guo, J. Sudweeks, and T. A. Dingus, "An analysis of driver inattention using a case-crossover approach on 100-car data: Final report," report, 2010.

[14] J. Tison, N. Chaudhary, and L. Cosgrove, "National phone survey on distracted driving attitudes and behaviors," report, 2011.

[15] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3, pp. 32–36, IEEE.

[16] K. Soomro and A. R. Zamir, *Action recognition in realistic sports videos*, pp. 181–208. Springer, 2014.

[17] M. Rodriguez, "Spatio-temporal maximum average correlation height templates in action recognition and video summarization," 2010.

[18] A. Bux, P. Angelov, and Z. Habib, *Vision based human activity recognition: a review*, pp. 341–371. Springer, 2017.

[19] A. B. Sargano, P. Angelov, and Z. Habib, "A comprehensive review on handcrafted and learning-based action representation approaches for human activity recognition," *Applied Sciences*, vol. 7, no. 1, p. 110, 2017.

[20] A. B. Sargano, P. Angelov, and Z. Habib, "Human action recognition from multiple views based on view-invariant feature descriptor using support vector machines," *Applied Sciences*, vol. 6, no. 10, p. 309, 2016.

[21] A. B. Sargano, X. Wang, P. Angelov, and Z. Habib, "Human action recognition using transfer learning with deep representations," in *Neural Networks (IJCNN), 2017 International Joint Conference on*, pp. 463–469, IEEE, 2017.

[22] O. Yurur, C. Liu, and W. Moreno, "A survey of context-aware middleware designs for human activity recognition," *IEEE Communications Magazine*, vol. 52, no. 6, pp. 24–31, 2014.

[23] S. Ranasinghe, F. Al Machot, and H. C. Mayr, "A review on applications of activity recognition systems with regard to performance and evaluation," *International Journal of Distributed Sensor Networks*, vol. 12, no. 8, p. 1550147716665520, 2016.

[24] T. Sztyler, H. Stuckenschmidt, and W. Petrich, "Position-aware activity recognition with wearable devices," *Pervasive and mobile computing*, 2017.

[25] H. Xu, J. Liu, H. Hu, and Y. Zhang, "Wearable sensor-based human activity recognition method with multi-features extracted from hilbert-huang transform," *Sensors*, vol. 16, no. 12, p. 2048, 2016.

[26] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu, "Sensor-based activity recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 790–808, 2012.

[27] T. Bouwmans, "Traditional and recent approaches in background modeling for foreground detection: An overview," *Computer Science Review*, vol. 11, pp. 31–66, 2014.

[28] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, 2013.

[29] M. Ramanathan, W.-Y. Yau, and E. K. Teoh, "Human action recognition with video data: research and evaluation challenges," *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 5, pp. 650–663, 2014.

[30] R. Poppe, "A survey on vision-based human action recognition," *Image and vision computing*, vol. 28, no. 6, pp. 976–990, 2010.

[31] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer vision and image understanding*, vol. 115, no. 2, pp. 224–241, 2011.

[32] M. Ziaeefard and R. Bergevin, "Semantic human activity recognition: a literature review," *Pattern Recognition*, vol. 48, no. 8, pp. 2329–2345, 2015.

[33] E. Maravelakis, A. Konstantaras, J. Kilty, E. Karapidakis, and E. Katsifarakis, "Automatic building identification and features extraction from aerial images: Application on the historic 1866 square of chania greece," in *Fundamentals of Electrical Engineering (ISFEE), 2014 International Symposium on*, pp. 1–6, IEEE.

[34] A. Jalal, S. Kamal, and D. Kim, "A depth video sensor-based life-logging human activity recognition system for elderly care in smart indoor environments," *Sensors*, vol. 14, no. 7, pp. 11735–11759, 2014.

[35] A. Jalal, N. Sarif, J. T. Kim, and T.-S. Kim, "Human activity recognition via recognized body parts of human depth silhouettes for residents monitoring services at smart home," *Indoor and Built Environment*, vol. 22, no. 1, pp. 271–279, 2013.

[36] J. Li and N. Allinson, "Building recognition using local oriented features," *IEEE Transactions on Industrial Informatics*, vol. 9, no. 3, pp. 1697–1704, 2013.

[37] A. Jalal, S. Kamal, and D. Kim, "Shape and motion features approach for activity tracking and recognition from kinect video camera," in *Advanced Information Networking and Applications Workshops (WAINA), 2015 IEEE 29th International Conference on*, pp. 445–450, IEEE.

[38] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[39] R. Yuan and W. Hui, "Object identification and recognition using multiple contours based moment invariants," in *2008 International Symposium on Information Science and Engineering*, vol. 1, pp. 140–144, IEEE.

[40] A. Jalal and Y. A. Rasheed, "Collaboration achievement along with performance maintenance in video streaming," in *Proceedings of the IEEE Conference on Interactive Computer Aided Learning, Villach, Austria*, vol. 2628, p. 18.

[41] S. KAMAL, C. A. AZURDIA-MEZA, and K. LEE, "Subsiding oob emission and ici power using ipower pulse in ofdm systems,"

[42] A. Farooq, A. Jalal, and S. Kamal, "Dense rgb-d map-based human tracking and activity recognition using skin joints features and self-organizing map," *KSII Transactions on internet and information systems*, vol. 9, no. 5, pp. 1856–1869, 2015.

[43] A. Jalal and S. Kim, "The mechanism of edge detection using the block matching criteria for the motion estimation," in *Proceedings of the Conference on Human Computer Interaction, Daegu, Korea*, vol. 3031, p. 484489.

[44] C. A. Azurdia-Meza, A. Falchetti, H. F. Arrano, S. Kamal, and K. Lee, "Evaluation of the improved parametric linear combination pulse in digital baseband communication systems," in *Information and Communication Technology Convergence (ICTC), 2015 International Conference on*, pp. 485–487, IEEE.

[45] S. Kamal and A. Jalal, "A hybrid feature extraction approach for human detection, tracking and activity recognition using depth sensors," *Arabian Journal for Science and Engineering*, vol. 41, no. 3, pp. 1043–1051, 2016.

[46] P. Bongale, A. Ranjan, and S. Anand, "Implementation of 3d object recognition and tracking," in *Recent Advances in Computing and Software Systems (RACSS), 2012 International Conference on*, pp. 77–79, IEEE.

[47] S. Kamal, A. Jalal, and D. Kim, "Depth images-based human detection, tracking and activity recognition using spatiotemporal features and modified hmm,"

[48] K. Lai, L. Bo, X. Ren, and D. Fox, "Sparse distance learning for object recognition combining rgb and depth information," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pp. 4007–4013, IEEE.

[49] A. Jalal, J. T. Kim, and T.-S. Kim, "Development of a life logging system via depth imaging-based human activity recognition for smart homes," in *Proceedings of the International Symposium on Sustainable Healthy Buildings, Seoul, Korea*, pp. 91–95.

[50] J.-Y. Chang, J.-J. Shyu, and C.-W. Cho, "Fuzzy rule inference based human activity recognition," in *2009 IEEE Control Applications,(CCA) Intelligent Control,(ISIC)*, pp. 211–215, IEEE.

[51] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, "Human pose estimation and activity recognition from multi-view videos: Comparative explorations of recent developments," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 6, no. 5, pp. 538–552, 2012.

[52] R. Gade and T. B. Moeslund, "Thermal cameras and applications: A survey," *Machine vision and applications*, vol. 25, no. 1, pp. 245–262, 2014.

[53] J. Han and B. Bhanu, "Human activity recognition in thermal infrared imagery," in *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pp. 17–17, IEEE, 2005.

[54] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, no. 3, p. 27, 2011.

[55] D. D. Dawn and S. H. Shaikh, "A comprehensive survey of human action recognition with spatio-temporal interest point (stip) detector," *The Visual Computer*, vol. 32, no. 3, pp. 289–306, 2016.

[56] I. Sipiran and B. Bustos, "Harris 3d: a robust extension of the harris operator for interest point detection on 3d meshes," *The Visual Computer*, vol. 27, no. 11, pp. 963–976, 2011.

[57] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[58] A. Gilbert, J. Illingworth, and R. Bowden, "Scale invariant action recognition using compound features mined from dense spatio-temporal corners," in *European Conference on Computer Vision*, pp. 222–233, Springer.

[59] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[60] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. S. Huang, "Action detection in complex scenes with spatial and temporal ambiguities," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 128–135, IEEE.

[61] M.-C. Roh, H.-K. Shin, and S.-W. Lee, "View-independent human action recognition with volume motion template on single stereo camera," *Pattern Recognition Letters*, vol. 31, no. 7, pp. 639–647, 2010.

[62] H. Qian, Y. Mao, W. Xiang, and Z. Wang, "Recognition of human activities using svm multi-class classifier," *Pattern Recognition Letters*, vol. 31, no. 2, pp. 100–111, 2010.

[63] X. Wu, D. Xu, L. Duan, and J. Luo, "Action recognition using context and appearance distribution features," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 489–496, IEEE.

[64] N. Ikizler and P. Duygulu, "Histogram of oriented rectangles: A new pose descriptor for human action recognition," *Image and Vision Computing*, vol. 27, no. 10, pp. 1515–1526, 2009.

[65] X. Peng, Y. Qiao, Q. Peng, and X. Qi, "Exploring motion boundary based sampling and spatial-temporal context descriptors for action recognition," in *BMVC*.

[66] J. Liu, B. Kuipers, and S. Savarese, "Recognizing human actions by attributes," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 3337–3344, IEEE.

[67] M. Chen, L. Gong, T. Wang, and Q. Feng, "Action recognition using lie algebrized gaussians over dense local spatio-temporal features," *Multimedia Tools and Applications*, vol. 74, no. 6, pp. 2127–2142, 2015.

[68] H. Wang, A. Klser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 3169–3176, IEEE.

[69] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 1234–1241, IEEE.

[70] S. Ma, L. Sigal, and S. Sclaroff, "Space-time tree ensemble for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5024–5032.

[71] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 915–922.

[72] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: a large video database for human motion recognition," in *2011 International Conference on Computer Vision*, pp. 2556–2563, IEEE.

[73] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3551–3558.

[74] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo, "Trajectory-based modeling of human actions with motion reference points," in *European Conference on Computer Vision*, pp. 425–438, Springer.

[75] O. Kliper-Gross, Y. Gurovich, T. Hassner, and L. Wolf, "Motion interchange patterns for action recognition in unconstrained videos," in *European Conference on Computer Vision*, pp. 256–269, Springer.

[76] L. Wang, Y. Qiao, and X. Tang, "Motionlets: Mid-level 3d parts for human motion recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2674–2681.

[77] X. Peng, C. Zou, Y. Qiao, and Q. Peng, "Action recognition with stacked fisher vectors," in *European Conference on Computer Vision*, pp. 581–595, Springer.

[78] M. Jain, H. Jegou, and P. Bouthemy, "Better exploiting motion for better action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2555–2562.

[79] B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati, and T. Tuytelaars, "Modeling video evolution for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5378–5387.

[80] M. Hoai and A. Zisserman, "Improving human action recognition using score distribution and ranking," in *Asian Conference on Computer Vision*, pp. 3–20, Springer.

[81] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2929–2936, IEEE.

[82] E. Vig, M. Dorr, and D. Cox, "Space-variant descriptor sampling for action recognition based on saliency and eye movements," in *European conference on computer vision*, pp. 84–97, Springer.

[83] S. Mathe and C. Sminchisescu, *Dynamic eye movement datasets and learnt saliency models for visual action recognition*, pp. 842–856. Springer, 2012.

[84] O. Kihl, D. Picard, and P.-H. Gosselin, "Local polynomial space–time descriptors for action classification," *Machine Vision and Applications*, vol. 27, no. 3, pp. 351–361, 2016.

[85] T. Lan, Y. Zhu, A. Roshan Zamir, and S. Savarese, "Action recognition by hierarchical mid-level action elements," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4552–4560.

[86] J. Yuan, Z. Liu, and Y. Wu, "Discriminative subvolume search for efficient action detection," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2442–2449, IEEE.

[87] B. B. Amor, J. Su, and A. Srivastava, "Action recognition using rate-invariant analysis of skeletal shape trajectories," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 1–13, 2016.

[88] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2752–2759.

[89] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 1996–2003, IEEE.

[90] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 984–989, IEEE.

[91] Y. Sheikh, M. Sheikh, and M. Shah, "Exploring the space of a human action," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 1, pp. 144–149, IEEE.

[92] J. Yang, Z. Shi, and Z. Wu, "Vision-based action recognition of construction workers using dense trajectories," *Advanced Engineering Informatics*, vol. 30, no. 3, pp. 327–336, 2016.

[93] Y.-G. Jiang, Q. Dai, W. Liu, X. Xue, and C.-W. Ngo, "Human action recognition in unconstrained videos by explicit motion modeling," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3781–3795, 2015.

[94] P. Dollr, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 65–72, IEEE.

[95] T. H. Thi, J. Zhang, L. Cheng, L. Wang, and S. Satoh, "Human action recognition and localization in video using structured learning of local space-time features," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, pp. 204–211, IEEE.

[96] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 1470–1477, IEEE.

[97] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Computer Vision and Image Understanding*, 2016.

[98] L. Liu, L. Wang, and X. Liu, "In defense of soft-assignment coding," in *2011 International Conference on Computer Vision*, pp. 2486–2493, IEEE.

[99] F. Perronnin, J. Snchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European conference on computer vision*, pp. 143–156, Springer.

[100] H. Wang, A. Klser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, vol. 103, no. 1, pp. 60–79, 2013.

[101] H. Li and M. Greenspan, "Multi-scale gesture recognition from time-varying contours," in *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, vol. 1, pp. 236–243, IEEE.

[102] C. Thurau and V. Hlavc, "Pose primitive based human action recognition in videos or still images," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE.

[103] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 726–733, IEEE.

[104] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE.

[105] Z. Jiang, Z. Lin, and L. Davis, "Recognizing human actions by learning and matching shape-motion prototype trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 533–547, 2012.

[106] M. B. Holte, T. B. Moeslund, N. Nikolaidis, and I. Pitas, "3d human action recognition for multi-view camera systems," in *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2011 International Conference on*, pp. 342–349, IEEE.

[107] P. Huang, A. Hilton, and J. Starck, "Shape similarity for 3d video sequences of people," *International Journal of Computer Vision*, vol. 89, no. 2-3, pp. 362–381, 2010.

[108] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 249–257, 2006.

[109] R. Slama, H. Wannous, M. Daoudi, and A. Srivastava, "Accurate 3d action recognition using learning on the grassmann manifold," *Pattern Recognition*, vol. 48, no. 2, pp. 556–567, 2015.

[110] L. Wang and D. Suter, "Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE.

[111] S. A. Rahman, S.-Y. Cho, and M. K. Leung, "Recognising human actions by analysing negative spaces," *Computer Vision, IET*, vol. 6, no. 3, pp. 197–213, 2012.

[112] D. Vishwakarma and R. Kapoor, "Hybrid classifier based human activity recognition using the silhouette and cells," *Expert Systems with Applications*, vol. 42, no. 20, pp. 6957–6965, 2015.

[113] I. N. Junejo, K. N. Junejo, and Z. Al Aghbari, "Silhouette-based human action recognition using sax-shapes," *The Visual Computer*, vol. 30, no. 3, pp. 259–269, 2014.

[114] A. A. Chaaraoui, P. Climent-Prez, and F. Flrez-Revuelta, "Silhouette-based human action recognition using sequences of key poses," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1799–1807, 2013.

[115] A. A. Chaaraoui and F. Flrez-Revuelta, "A low-dimensional radial silhouette-based feature for fast human action recognition fusing multiple views," *International Scholarly Research Notices*, vol. 2014, 2014.

[116] S. A. Rahman, I. Song, M. K. Leung, I. Lee, and K. Lee, "Fast action recognition using negative space features," *Expert Systems with Applications*, vol. 41, no. 2, pp. 574–587, 2014.

[117] S. Cheema, A. Eweiwi, C. Thurau, and C. Bauckhage, "Action recognition by learning discriminative key poses," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pp. 1302–1309, IEEE.

[118] S. Chun and C.-S. Lee, "Human action recognition using histogram of motion intensity and direction from multiple views," *IET Computer Vision*, 2016.

[119] F. Murtaza, M. H. Yousaf, and S. Velastin, "Multi-view human action recognition using 2d motion templates based on mhis and their hog description," *IET Computer Vision*, 2016.

[120] M. Ahmad and S.-W. Lee, "Hmm-based human action recognition using multiview image sequences," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1, pp. 263–266, IEEE.

[121] D. K. Vishwakarma, R. Kapoor, and A. Dhiman, "A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics," *Robotics and Autonomous Systems*, vol. 77, pp. 25–38, 2016.

[122] S. Pehlivan and D. A. Forsyth, "Recognizing activities in multiple views with fusion of frame judgments," *Image and Vision Computing*, vol. 32, no. 4, pp. 237–249, 2014.

[123] A. Eweiwi, S. Cheema, C. Thurau, and C. Bauckhage, "Temporal key poses for human action recognition," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pp. 1310–1317, IEEE.

[124] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *Pattern Recognition, 1994. Vol. 1-Conference A: Computer Vision amp; Image*

Processing., Proceedings of the 12th IAPR International Conference on, vol. 1, pp. 582–585, IEEE.

[125] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[126] M. Pietikinen, A. Hadid, G. Zhao, and T. Ahonen, *Computer vision using local binary patterns*, vol. 40. Springer Science Business Media, 2011.

[127] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[128] L. Yeffet and L. Wolf, "Local trinary patterns for human action recognition," in *2009 IEEE 12th International Conference on Computer Vision*, pp. 492–497, IEEE.

[129] V. Kellokumpu, G. Zhao, and M. Pietikinen, "Human activity recognition using a dynamic texture based method," in *BMVC*, vol. 1, p. 2.

[130] A. K. S. Kushwaha, S. Srivastava, and R. Srivastava, "Multi-view human activity recognition based on silhouette and uniform rotation invariant local binary patterns," *Multimedia Systems*, pp. 1–17, 2016.

[131] F. Baumann, A. Ehlers, B. Rosenhahn, and J. Liao, "Recognizing human actions using novel space-time volume binary patterns," *Neurocomputing*, vol. 173, pp. 54–63, 2016.

[132] S. Sadek, A. Al-Hamadi, B. Michaelis, and U. Sayed, "An action recognition scheme using fuzzy log-polar histogram and temporal self-similarity," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 1–9, 2011.

[133] B. Yao, H. Hagras, M. J. Alhaddad, and D. Alghazzawi, "A fuzzy logic-based system for the automation of human behavior recognition using machine vision in intelligent environments," *Soft Computing*, vol. 19, no. 2, pp. 499–506, 2015.

[134] C. H. Lim and C. S. Chan, "Fuzzy qualitative human model for viewpoint identification," *Neural Computing and Applications*, vol. 27, no. 4, pp. 845–856, 2016.

[135] T. Obo, C. K. Loo, M. Seera, and N. Kubota, "Hybrid evolutionary neuro-fuzzy approach based on mutual adaptation for human gesture recognition," *Applied Soft Computing*, vol. 42, pp. 377–389, 2016.

[136] B. Yousefi and C. K. Loo, "Bio-inspired human action recognition using hybrid max-product neuro-fuzzy classifier and quantum-behaved pso," *arXiv preprint arXiv:1509.03789*, 2015.

[137] J. A. Iglesias, P. Angelov, A. Ledezma, and A. Sanchis, "Evolving classification of agents behaviors: a general approach," *Evolving Systems*, vol. 1, no. 3, pp. 161–171, 2010.

[138] V. Kellokumpu, G. Zhao, and M. Pietikinen, "Recognition of human actions using texture descriptors," *Machine Vision and Applications*, vol. 22, no. 5, pp. 767–780, 2011.

[139] S. Sadek, A. Al-Hamadi, B. Michaelis, and U. Sayed, "Human action recognition via affine moment invariants," in *Pattern Recognition (ICPR), 2012 21st International Conference on*, pp. 218–221, IEEE.

[140] R. Mattivi and L. Shao, "Human action recognition using lbp-top as sparse spatio-temporal feature descriptor," in *International Conference on Computer Analysis of Images and Patterns*, pp. 740–747, Springer.

[141] D. Weinland, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3d exemplars," in *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–7, IEEE.

[142] Z. Lin, Z. Jiang, and L. S. Davis, "Recognizing actions by shape-motion prototype trees," in *2009 IEEE 12th international conference on computer vision*, pp. 444–451, IEEE.

[143] M. B. Holte, B. Chakraborty, J. Gonzalez, and T. B. Moeslund, "A local 3-d motion descriptor for multi-view human action recognition from 4-d spatio-temporal interest points," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 5, pp. 553–565, 2012.

[144] P. Turaga, A. Veeraraghavan, and R. Chellappa, "Statistical analysis on stiefel and grassmann manifolds with applications in computer vision," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE.

[145] S. Pehlivan and P. Duygulu, "A new pose-based representation for recognizing actions from multiple cameras," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 140–151, 2011.

[146] F. Zhu, L. Shao, J. Xie, and Y. Fang, "From handcrafted to learned representations for human action recognition: A survey," *Image and Vision Computing*, 2016.

[147] T. Guha and R. K. Ward, "Learning sparse representations for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1576–1588, 2012.

[148] H. Wang, C. Yuan, W. Hu, and C. Sun, "Supervised class-specific dictionary learning for sparse modeling in action recognition," *Pattern Recognition*, vol. 45, no. 11, pp. 3902–3911, 2012.

[149] J. Zheng, Z. Jiang, P. J. Phillips, and R. Chellappa, "Cross-view action recognition via a transferable dictionary pair," in *BMVC*, vol. 1, p. 7.

[150] J. Zheng, Z. Jiang, and R. Chellappa, "Cross-view action recognition via transferable dictionary learning," 2016.

[151] F. Zhu and L. Shao, "Weakly-supervised cross-domain dictionary learning for visual recognition," *International Journal of Computer Vision*, vol. 109, no. 1-2, pp. 42–59, 2014.

[152] F. Zhu and L. Shao, "Correspondence-free dictionary learning for cross-view action recognition," in *ICPR*, pp. 4525–4530.

[153] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 3360–3367, IEEE.

[154] L. Liu, L. Shao, X. Li, and K. Lu, "Learning spatio-temporal representations for action recognition: a genetic programming approach," *IEEE transactions on cybernetics*, vol. 46, no. 1, pp. 158–170, 2016.

[155] Y. Zhang, Y. Zhang, E. Swears, N. Larios, Z. Wang, and Q. Ji, "Modeling temporal interactions with interval temporal bayesian networks for complex activity recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 10, pp. 2468–2483, 2013.

[156] F. M. Khan, S. C. Lee, and R. Nevatia, "Conditional bayesian networks for action detection," in *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on*, pp. 256–262, IEEE, 2013.

[157] C. Y. Park, K. B. Laskey, P. C. Costa, and S. Matsumoto, "A process for human-aided multi-entity bayesian networks learning in predictive situation awareness," in *Information Fusion (FUSION), 2016 19th International Conference on*, pp. 2116–2124, IEEE, 2016.

[158] L. Deng and D. Yu, "Deep learning," *Signal Processing*, vol. 7, pp. 3–4, 2014.

[159] A. Ivakhnenko, "Polynomial theory of complex systems," *IEEE Transactions on Systems, Man, and Cybernetics*, no. 4, pp. 364–378, 1971.

[160] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[161] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[162] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[163] P. Smolensky, "Information processing in dynamical systems: Foundations of harmony theory," report, DTIC Document, 1986.

[164] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 3361–3368, IEEE.

[165] P. Foggia, A. Saggese, N. Strisciuglio, and M. Vento, "Exploiting the deep learning paradigm for recognizing human actions," in *Advanced Video and Signal Based Surveillance (AVSS), 2014 11th IEEE International Conference on*, pp. 93–98, IEEE.

[166] M. Hasan and A. K. Roy-Chowdhury, "Continuous learning of human activity models using deep nets," in *European Conference on Computer Vision*, pp. 705–720, Springer.

[167] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra, "Effective codebooks for human action representation and classification in unconstrained videos," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 1234–1245, 2012.

[168] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105.

[169] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[170] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision*, pp. 818–833, Springer.

[171] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *2011 International Conference on Computer Vision*, pp. 2018–2025, IEEE.

[172] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 1725–1732.

[173] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[174] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural computation*, vol. 14, no. 4, pp. 715–770, 2002.

[175] Z. Zhang and D. Tao, "Slow feature analysis for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 3, pp. 436–450, 2012.

[176] L. Sun, K. Jia, T.-H. Chan, Y. Fang, G. Wang, and S. Yan, "Dl-sfa: deeply-learned slow feature analysis for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2625–2632.

[177] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1110–1118.

[178] L. Sun, K. Jia, D.-Y. Yeung, and B. E. Shi, "Human action recognition using factorized spatio-temporal convolutional networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4597–4605.

[179] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 4489–4497, IEEE.

[180] E. Park, X. Han, T. L. Berg, and A. C. Berg, "Combining multiple sources of knowledge in deep cnns for action recognition," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–8, IEEE.

[181] S. Yu, Y. Cheng, S. Su, G. Cai, and S. Li, "Stratified pooling based deep convolutional neural networks for human action recognition," *Multimedia Tools and Applications*, pp. 1–16, 2016.

[182] E. P. Ijjina and C. K. Mohan, "Human action recognition based on motion capture information using fuzzy convolution neural networks," in *Advances in Pattern Recognition (ICAPR), 2015 Eighth International Conference on*, pp. 1–6, IEEE, 2015.

[183] G. Chron, I. Laptev, and C. Schmid, "P-cnn: Pose-based cnn features for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3218–3226.

[184] G. Gkioxari, R. Girshick, and J. Malik, "Contextual action recognition with r* cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1080–1088.

[185] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587.

[186] H. Rahmani and A. Mian, "3d action recognition from novel viewpoints," in *CVPR, June*.

[187] A. Alfaro, D. Mery, and A. Soto, "Action recognition in video using sparse coding and relative features," *arXiv preprint arXiv:1605.03222*, 2016.

[188] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[189] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4305–4314.

[190] Z. Lan, M. Lin, X. Li, A. G. Hauptmann, and B. Raj, "Beyond gaussian pyramid: Multi-skip feature stacking for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 204–212.

[191] H. Bilen, B. Fernando, E. Gavves, A. Vedaldi, and S. Gould, "Dynamic image networks for action recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition CVPR*.

[192] B. Mahasseni and S. Todorovic, "Regularizing long short term memory with 3d human-skeleton sequences for action recognition,"

[193] B. Fernando, E. Gavves, J. Oramas, A. Ghodrati, and T. Tuytelaars, "Rank pooling for action recognition," 2016.

[194] W. Zhu, J. Hu, G. Sun, X. Cao, and Y. Qiao, "A key volume mining deep framework for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1991–1999.

[195] C. Wang, Y. Wang, and A. L. Yuille, "Mining 3d key-pose-motifs for action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2639–2647.

[196] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4041–4049.

[197] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4694–4702.

[198] P. Weinzaepfel, Z. Harchaoui, and C. Schmid, "Learning to track for spatio-temporal action localization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3164–3172.

[199] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 961–970.

[200] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pp. 2929–2936, IEEE, 2009.

[201] K. K. Reddy and M. Shah, "Recognizing 50 human action categories of web videos," *Machine Vision and Applications*, vol. 24, no. 5, pp. 971–981, 2013.

[202] D. Rudoy and L. Zelnik-Manor, "Viewpoint selection for human actions," *International Journal of Computer Vision*, vol. 97, no. 3, pp. 243–254, 2012.

[203] B. Saghafi, D. Rajan, and W. Li, "Efficient 2d viewpoint combination for human action recognition," *Pattern Analysis and Applications*, vol. 19, no. 2, pp. 563–577, 2016.

[204] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 2, pp. 288–303, 2010.

[205] M. B. Holte, C. Tran, M. M. Trivedi, and T. B. Moeslund, "Human action recognition using multiple views: a comparative perspective on recent developments," in *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, pp. 47–52, ACM.

[206] I. N. Junejo, E. Dexter, I. Laptev, and P. Perez, "View-independent action recognition from temporal self-similarities," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 1, pp. 172–185, 2011.

[207] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 23, no. 3, pp. 412–424, 2012.

[208] A. Iosifidis, A. Tefas, and I. Pitas, "Multi-view action recognition based on action volumes, fuzzy distances and cluster discriminant analysis," *Signal Processing*, vol. 93, no. 6, pp. 1445–1457, 2013.

[209] J. Lei, G. Li, J. Zhang, Q. Guo, and D. Tu, "Continuous action segmentation and recognition using hybrid convolutional neural network-hidden markov model model," *IET Computer Vision*, 2016.

[210] N. Gkalelis, N. Nikolaidis, and I. Pitas, "View indepedent human movement recognition from multi-view video exploiting a circular invariant posture representation," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pp. 394–397, IEEE.

[211] D. Weinland, M. zuysal, and P. Fua, *Making action recognition robust to occlusions and viewpoint changes*, pp. 635–648. Springer, 2010.

[212] R. Souvenir and J. Babbs, "Learning the viewpoint manifold for action recognition," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–7, IEEE.

[213] J. Liu and M. Shah, "Learning human actions via information maximization," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE.

[214] W. Nie, A. Liu, W. Li, and Y. Su, "Cross-view action recognition by cross-domain learning," *Image and Vision Computing*, 2016.

[215] C.-H. Hsieh, P. S. Huang, and M.-D. Tang, "Human action recognition using silhouette histogram," in *Proceedings of the Thirty-Fourth Australasian Computer Science Conference-Volume 113*, pp. 11–16, Australian Computer Society, Inc.

[216] M.-K. Hu, "Visual pattern recognition by moment invariants," *information Theory, IRE Transactions on*, vol. 8, no. 2, pp. 179–187, 1962.

[217] Z. Huang and J. Leng, "Analysis of hu's moment invariants on image scaling and rotation," in *Computer Engineering and Technology (ICCET), 2010 2nd International Conference on*, vol. 7, pp. V7–476, IEEE, 2010.

[218] A. Jordan, "On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes," *Advances in neural information processing systems*, vol. 14, p. 841, 2002.

[219] V. N. Vapnik and V. Vapnik, *Statistical learning theory*, vol. 1. Wiley New York, 1998.

[220] U. H.-G. Kreel, "Pairwise classification and support vector machines," in *Advances in kernel methods*, pp. 255–268, MIT Press.

[221] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin dags for multiclass classification," in *nips*, vol. 12, pp. 547–553.

[222] T. G. Dietterich and G. Bakiri, "Solving multiclass learning problems via error-correcting output codes," *Journal of artificial intelligence research*, pp. 263–286, 1995.

[223] S. Cheong, S. H. Oh, and S.-Y. Lee, "Support vector machines with binary tree architecture for multi-class classification," *Neural Information Processing-Letters and Reviews*, vol. 2, no. 3, pp. 47–51, 2004.

[224] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *Neural Networks, IEEE Transactions on*, vol. 13, no. 2, pp. 415–425, 2002.

[225] K. Manosha Chathuramali and R. Rodrigo, "Faster human activity recognition with svm," in *Advances in ICT for Emerging Regions (ICTer), 2012 International Conference on*, pp. 197–203, IEEE.

[226] K. K. Reddy, J. Liu, and M. Shah, "Incremental action recognition using feature-tree," in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 1010–1017, IEEE.

[227] F. Lv and R. Nevatia, "Single view human action recognition using key pose matching and viterbi path searching," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8, IEEE.

[228] S. N. Vitaladevuni, V. Kellokumpu, and L. S. Davis, "Action recognition using ballistic dynamics," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE.

[229] S. Cherla, K. Kulkarni, A. Kale, and V. Ramasubramanian, "Towards fast, view-invariant human action recognition," in *Computer Vision and Pattern Recognition Workshops, 2008. CVPRW'08. IEEE Computer Society Conference on*, pp. 1–8, IEEE.

[230] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[231] X. Cao, Z. Wang, P. Yan, and X. Li, "Transfer learning for pedestrian detection," *Neurocomputing*, vol. 100, pp. 51–57, 2013.

[232] D. Wu, F. Zhu, and L. Shao, "One shot learning gesture recognition from rgbd images," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pp. 7–12, IEEE, 2012.

[233] L. Fei-Fei, "Knowledge transfer in learning to recognize visual objects classes," in *International Conference on Development and Learning*, pp. 1–8, 2006.

[234] H. Azizpour, A. Sharif Razavian, J. Sullivan, A. Maki, and S. Carlsson, "From generic to specific deep representations for visual recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 36–45.

[235] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[236] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1717–1724.

[237] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," *arXiv preprint arXiv:1510.07945*, 2015.

[238] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1520–1528.

[239] L. Shao, F. Zhu, and X. Li, "Transfer learning for visual categorization: A survey," *IEEE transactions on neural networks and learning systems*, vol. 26, no. 5, pp. 1019–1034, 2015.

[240] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 806–813.

[241] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–9.

[242] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," *Computer Vision–ECCV 2008*, pp. 650–663, 2008.

[243] A. Klaser, M. Marszaek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC 2008-19th British Machine Vision Conference*, pp. 275: 1–10, British Machine Vision Association.

[244] A. Fathi, Y. Li, and J. M. Rehg, "Learning to recognize daily actions using gaze," in *European Conference on Computer Vision*, pp. 314–327, Springer, 2012.

[245] C. Yuan, X. Li, W. Hu, H. Ling, and S. J. Maybank, "Modeling geometric-temporal context with directional pyramid co-occurrence for action recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 658–672, 2014.

[246] P. Angelov and A. Sperduti, "Challenges in deep learning," in *Proc. European Symp. on Artificial NNs*, pp. 485–495.

[247] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *European conference on computer vision*, pp. 140–153, Springer.

[248] K. Charalampous and A. Gasteratos, "On-line deep learning method for action recognition," *Pattern Analysis and Applications*, vol. 19, no. 2, pp. 337–354, 2016.

[249] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[250] L. Cao, Z. Liu, and T. S. Huang, "Cross-dataset action detection," in *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pp. 1998–2005, IEEE.

[251] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pp. 321–330, ACM, 2006.

[252] Y. Aytar, *Transfer learning for object category detection.* Thesis, 2014.

[253] Y.-C. Su, T.-H. Chiu, C.-Y. Yeh, H.-F. Huang, and W. H. Hsu, "Transfer learning for video recognition with scarce training data for deep convolutional neural network," *arXiv preprint arXiv:1409.4127*, 2014.

[254] M. A. R. Ahad, M. N. Islam, and I. Jahan, "Action recognition based on binary patterns of action-history and histogram of oriented gradient," *Journal on Multimodal User Interfaces*, vol. 10, no. 4, pp. 335–344, 2016.

[255] S. Ding and S. Qu, "An improved interest point detector for human action recognition," in *Control and Decision Conference (CCDC), 2016 Chinese*, pp. 4355–4360, IEEE, 2016.

[256] Y. Tian, Q. Ruan, G. An, and Y. Fu, "Action recognition using local consistent group sparse coding with spatio-temporal structure," in *Proceedings of the 2016 ACM on Multimedia Conference*, pp. 317–321, ACM.

[257] I. Atmosukarto, N. Ahuja, and B. Ghanem, "Action recognition using discriminative structured trajectory groups," in *2015 IEEE Winter Conference on Applications of Computer Vision*, pp. 899–906, IEEE.

[258] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pp. 2046–2053, IEEE.

[259] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC 2009-British Machine Vision Conference*, pp. 124.1–124.11, BMVA Press.

[260] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pp. 1–8, IEEE, 2008.

[261] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *Proceedings of the 26th annual international conference on machine learning*, pp. 609–616, ACM, 2009.

[262] S. Fernández, A. Graves, and J. Schmidhuber, "Sequence labelling in structured domains with hierarchical recurrent neural networks.," in *IJCAI*, pp. 774–779, 2007.

[263] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*, pp. 6645–6649, IEEE, 2013.

[264] R. Salakhutdinov and G. Hinton, "Deep boltzmann machines," in *Artificial Intelligence and Statistics*, pp. 448–455, 2009.

[265] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988–999, 1999.

[266] S. Linnainmaa, "The representation of the cumulative rounding error of an algorithm as a taylor expansion of the local rounding errors," *Master's Thesis (in Finnish), Univ. Helsinki*, pp. 6–7, 1970.

[267] P. J. Werbos, "Applications of advances in nonlinear sensitivity analysis," in *System modeling and optimization*, pp. 762–770, Springer, 1982.

[268] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, pp. 1026–1034, 2015.

[269] Y. Xia, L. Zhang, W. Xu, Z. Shan, and Y. Liu, "Recognizing multi-view objects with occlusions using a deep architecture," *Information Sciences*, vol. 320, pp. 333–345, 2015.

[270] D. Cireşan, U. Meier, J. Masci, and J. Schmidhuber, "A committee of neural networks for traffic sign classification," in *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pp. 1918–1921, IEEE, 2011.

[271] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1891–1898, 2014.

[272] D. Turcsany and A. Bargiela, "Learning local receptive fields in deep belief networks for visual feature detection," in *International Conference on Neural Information Processing*, pp. 462–470, Springer, 2014.

[273] G. E. Hinton and T. J. Sejnowski, "Optimal perceptual inference," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 448–453, IEEE New York, 1983.

[274] P. Xie, Y. Deng, and E. Xing, "Diversifying restricted boltzmann machine for document modeling," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1315–1324, ACM, 2015.

[275] S. Elfwing, E. Uchibe, and K. Doya, "Expected energy-based restricted boltzmann machine for classification," *Neural networks*, vol. 64, pp. 29–38, 2015.

[276] S. A. Eslami, N. Heess, C. K. Williams, and J. Winn, "The shape boltzmann machine: a strong model of object shape," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 155–176, 2014.

[277] D. Turcsany, A. Bargiela, and T. Maul, "Local receptive field constrained deep networks," *Information Sciences*, vol. 349, pp. 229–247, 2016.

[278] S. Nie, Z. Wang, and Q. Ji, "A generative restricted boltzmann machine based method for high-dimensional motion data modeling," *Computer Vision and Image Understanding*, vol. 136, pp. 14–22, 2015.

[279] L. Xu, Y. Li, Y. Wang, and E. Chen, "Temporally adaptive restricted boltzmann machine for background modeling.," in *AAAI*, pp. 1938–1944, 2015.

[280] F. Zhao, Y. Huang, L. Wang, T. Xiang, and T. Tan, "Learning relevance restricted boltzmann machine for unstructured group activity and event understanding," *International Journal of Computer Vision*, vol. 119, no. 3, pp. 329–345, 2016.

[281] G. Dahl, A.-r. Mohamed, G. E. Hinton, *et al.*, "Phone recognition with the mean-covariance restricted boltzmann machine," in *Advances in neural information processing systems*, pp. 469–477, 2010.

[282] T. Maniak, C. Jayne, R. Iqbal, and F. Doctor, "Automated intelligent system for sound signalling device quality assurance," *Information Sciences*, vol. 294, pp. 600–611, 2015.

[283] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the national academy of sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[284] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural computation*, vol. 14, no. 8, pp. 1771–1800, 2002.

[285] G. E. Hinton, "Learning multiple layers of representation," *Trends in cognitive sciences*, vol. 11, no. 10, pp. 428–434, 2007.

[286] G. E. Hinton, P. Dayan, B. J. Frey, and R. M. Neal, "The" wake-sleep" algorithm for unsupervised neural networks," *Science*, vol. 268, no. 5214, p. 1158, 1995.

[287] Z. Ghahramani, "Bayesian non-parametrics and the probabilistic approach to modelling," *Phil. Trans. R. Soc. A*, vol. 371, no. 1984, p. 20110553, 2013.

[288] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data," *Machine learning*, vol. 20, no. 3, pp. 197–243, 1995.

[289] M. Pelikan, K. Sastry, and D. E. Goldberg, "Evolutionary algorithms+ graphical models= scalable black-box optimization," *IlliGAL report*, no. 2001029, 2001.

[290] P. P. Angelov and D. P. Filev, "Flexible models with evolving structure," *International Journal of Intelligent Systems*, vol. 19, no. 4, pp. 327–340, 2004.

[291] Y. Tian, R. Sukthankar, and M. Shah, "Spatiotemporal deformable part models for action detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2642–2649, 2013.

[292] A.-A. Liu, N. Xu, W.-Z. Nie, Y.-T. Su, Y. Wong, and M. Kankanhalli, "Benchmarking a multimodal and multiview and interactive dataset for human action recognition," *IEEE Transactions on cybernetics*, 2017.