Contributions into Holistic Human Action Recognition

By

Ignace Tchangou Toudjeu

A thesis submitted in fulfillment of the academic requirements for the Degree of Doctor of Philosophy in Computer Engineering in the School of Engineering University of KwaZulu-Natal Durban, South Africa

August 2020

UNIVERSITY OF KWAZULU-NATAL, COLLEGE OF AGRICULTURE, ENGINEERING AND SCIENCE DECLARATION

The research described in this thesis was performed at the University of KwaZulu-Natal under the supervision of Prof. Jules-Raymond Tapamo. I hereby declare that all materials incorporated in this thesis are my own original work except where acknowledgement is made by name or in form of reference. The work contained herein has not been submitted in part or whole for a degree at any other university.

Signed:.....

Ignace Tchangou Toudjeu

Date: August 2020

As the candidate's supervisor, I have approved this thesis for submission.

Signed:.....

Prof. Jules-Raymond Tapamo

Date: August 2020

UNIVERSITY OF KWAZULU-NATAL, COLLEGE OF AGRICULTURE, ENGINEERING AND SCIENCE DECLARATION 1 -PLAGIARISM

I, IGNACE TCHANGOU TOUDJEU, declare that:

1. The research reported in this thesis, except where otherwise indicated, is my original research.

2. This thesis has not been submitted for any degree or examination at any other university.

3. This thesis does not contain other persons data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

4. This thesis does not contain other persons writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then

   a. Their words have been re-written but the general information attributed to them has been referenced;

   b. Where their exact words have been used, then their writing has been placed inside quotation marks, and referenced.

5. Where I have reproduced a publication of which I am an author, co-author or editor, I have indicated in detail which part of the publication was actually written by myself alone and have fully referenced such publications.

6. This thesis does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the thesis and in the References sections.

Signed:  .....

UNIVERSITY OF KWAZULU-NATAL, COLLEGE OF AGRICULTURE, ENGINEERING AND SCIENCE DECLARATION 2 -PUBLICATIONS

I, IGNACE TCHANGOU TOUDJEU , declare that the following publications came out of this thesis

1. I. Tchangou Toudjeu and J.R. Tapamo, Slope Pattern Spectra for Human Action Recognition, *Lecture notes in Computer Science*, vol. 10882, pp. 381-389, June 2018.

2. I. Tchangou Toudjeu and J.R. Tapamo, Circular Derivative Local Binary Pattern Feature Description for Facial Expression Recognition." *Advances in Electrical and Computer Engineering*, vol. 19, no. 1, pp. 51-56, 2019.

3. I. Tchangou Toudjeu and J.R. Tapamo, A 2D Convolution Neural Network for Human Action Recognition. In 2019 IEEE AFRICON, pp. 1-5, September 2019.

4. I. Tchangou Toudjeu and J.R. Tapamo, Pedestrian Action Recognition using Circular Derivative Binary Patterns and Histogram of Oriented Gradient. Manuscript submitted in Pattern Recognition Letters.

Signed:.....

"Tis human actions paint the chart of time"

James Montgemery

Abstract

In this thesis we holistically investigate the interpretation of human actions in both still images and videos. Human action recognition is currently a research problem of great interest both in academia and industry due to its potential applications which include security surveillances, sports annotation, human-computer interaction, and robotics. Action recognition, being a process of labelling actions using sensory observations, can be defined as a sequence of movements engendered by a human during an executed task. Such a process, when considering visual observations, is quite challenging and faces issues such as background clutter, shadows, illumination variations, occlusions, changes in scale, changes in the person performing the action, and viewpoint variations. Although many approaches to development of human action recognition systems have been proposed in the literature, they focused more on recognition accuracy while ignoring the computational complexity accompanying the recognition process. However, a human action recognition system which is both effective and efficient and can be operated real-time is needed. Firstly, we review, evaluate and compare the most prominent state-of-the-art feature extraction representations categorized between handcrafted feature based techniques and deep learning feature based techniques. Secondly, we propose holistic approaches in each of the categories. The first holistic approach takes advantage of existing slope patterns in the motion history images, which are a simple two dimensional representation of video, and reduces the running time of action recognition. The second one based on circular derivative local binary patterns outperforms the LBP based state-of-the-art techniques and addresses the issues of dimensionality by producing feature descriptor with minimal dimension size with less compromise on the recognition accuracy. The third one introduces a preprocessing step in a proposed 2D-convolutional neural network to deal with the same issue of dimensionality differently in the deep learning techniques. Here the temporal dimension is embedded into motion history images before being learned by a two dimensional convolutional neural network. Thirdly, three datasets (JAFFE, KTH and Pedestrian Action dataset) were used to validate the proposed human action recognition models. Finally, we show that better performance in comparison to the state-of-the-art methods can be achieved using holistic feature based techniques.

Key terms:

Human Action Recognition; Motion History Image; Circular Derivative Local Binary Pattern; Convolutional Neural Network; Facial Expression Recognition; Spatio-Temporal features

Acknowledgements

First of all, I would like to thank my supervisor Prof. Jules-Raymond Tapamo for guiding me well throughout this research work from the selection of the title to the finding of results. His immense knowledge, patience and motivation have equipped me with more power and spirit to excel in the research writing. Without his persistent help and companionship during sleepless nights, the goal of this research work would have not been realized. I am highly privileged to have worked under his supervision on such a challenging research topic in the field of computer vision. I thank you.

Apart from my Supervisor, I would like to thank Dr. Remy Tiako for his support and care during my PhD studies. He always made sure I did not deviate from my research goal. I thank the Postgraduate administrators, Ms. Ausie Luthuli and Ms. Nombuso Dlamini, for their administrative professionalism shown by their responsiveness at postgraduate student matters. My sincere gratitude also go to the University of KwaZulu-Natal for financial assistance, which included the sponsorship for an international trip.

Special thanks go to Prof. Barend Jacobus van Wyk who initiated me in academic research, never doubted on my research potential, and at crossroads encouraged me to further my studies. I thank Mr. Johann Pretorius for once telling me that I was young and should consider registereing for a PhD degree. I also thank Gustave Udahemuka for his everlasting friendship.

To my moral and emotional coaches, Mama Chantal and Mama Annie, I thank you and appreciate your reminders, compliments, and beliefs towards who I am and what I am capable of doing. I thank you both.

Finally, I wish to acknowledge the support and great love of my family, my wife, Annie; my kids, Béatrice and Ivan. They kept me going on and this work would not have been possible without their encouragement. A big thank to my parents and siblings for also affording me time and space during my studies and I hope for their understanding for not often being there. Without forgetting my source of strength, I thank my ancestors and God Almighty.

Contents

List of Figures

List of Tables

List of acronyms

AAM	**A**ctive **A**ppearance **M**odel
ASM	**A**ctive **S**hape **M**odel
BoW	**B**ag-**o**f-**W**ords
CD-LBP	**C**ircular **D**erivative-**L**ocal **B**inary **P**attern
CNN	**C**onvolution **N**eural **N**etwork
CPU	**C**entral **P**rocessor **U**nit
CS-LBP	**C**enter-**S**ymmetric **L**ocal **B**inary **P**attern
DBM	**D**eep **B**elief **M**achine
DBN	**D**eep **B**elief **N**etwork
DL	**D**eep-**L**earning
DNN	**D**eep **N**eural **N**etwork
FC	**F**ull- **C**onnected
FER	**F**acial **E**xpression **R**ecognition
FN	**F**alse **N**egative
FP	**F**alse **P**ositive
GPU	**G**raphical **P**rocessor **U**nit
HAR	**H**uman **A**ction **R**ecognition
HMM	**H**idden **M**arkov **M**odel
HOG	**H**istogram of **O**riented **G**radients
KNN	**K**-**N**earest **N**eighbor
LBP	**L**ocal **B**inary **P**attern
LBPV	**L**ocal **B**inary **P**attern **V**ariance
LTC	**L**ong-term **T**emporal-**M**emory
LTSM	**L**ong-**S**hort **T**erm **M**emory
MEI	**M**otion **E**nergy **I**mage

MHI	**M**otion **H**istory **I**mage
MHI-CDLBP-HOG	**M**otion **H**istory **I**mage-**C**ircular **D**erivative **L**ocal **B**inary **P**attern-**H**istogram of **O**riented **G**radients
MHI-SPS	**M**otion **H**istory **I**mage-**S**lope **P**attern **S**pectra
MHK	**M**HI-**H**OG-**K**NN
MHS	**M**HI-**H**OG-**S**VM
MHV	**M**otion **H**istory **V**olume
MLBP	**M**ultiscale **L**ocal **B**inary **P**attern
PAR	**P**edestrian **A**ction **R**ecognition
PCA	**P**rinciple **C**omponent **A**nalysis
ReLU	**Re**ctified **L**inear **U**nits
RNN	**R**ecurrent **N**eural **N**etwork
ROI	**R**egion **O**f **I**nterest
SIFT	**S**cale -**I**nvariant **F**eature **T**ransform
SPS	**S**lope **P**attern **S**pectra
SSS	**S**mall **S**ample **S**ize
STIP	**S**patio-**T**emporal **I**nterest **P**oint
SVM	**S**upport **V**ector **M**achines
TN	**T**rue **N**egative
TP	**T**rue **P**ositive

Dedicated to my family; my wife Annie and children Béatrice and Ivan. I love you all.

Chapter 1

Introduction

"Vision without action is merely a dream. Action without vision just passes the time. Vision with action can change the world."

Joel A. Barker

1.1 Introduction

The ability to recognize human actions is a demonstration of incredible human intelligence. Over the past two decades many researchers made attempts to mimic the remarkable visual perception of human beings in the field of computer vision. Within the human action recognition (HAR) literature, how human beings perceive human actions has become the center of an important and active research area. Many visual-based applications have been developed for the recognition of human actions. These applications, whose goal is to fully or partially observe human being in a still image or in a video sequence and identify automatically what they do as actions or portray as emotions, are found in health care [18][19][20], sports [21][22], video surveillance [23][24], human computer interaction [25][26][27], or robotics [28]. Human action recognition is also associated with facial expression recognition (FER) [29], which is often applied to still images. In this case, facial muscle activities are captured in still images resulting to emotional expressions characterized by facial actions [30]. Though HAR is currently a research problem of great interest both in academia and in industry due to its potential applications, many HAR systems still encountering some challenges have been developed. A common challenge to HAR in the literature is the development of an effective HAR system that matches the real human interpretation by overcoming problems such as occlusions, background clutter, changes in scale, illumination and appearance,

camera viewpoint, resolution of frames, large video data size, and noise [29]. Some representative frames highlighting some of the mentioned variabilities as challenges in a *boxing* action are depicted in Figure 1.1. Another obstacle to HAR is the personality in performing actions. Actions performed by different persons differ as well as actions that are repeated by the very same person. Moreover HAR task becomes even more complex when considering only still images due to the lack of temporal information. A video contains both spatial and temporal information, whereas an image only contains a spatial information. The type of information has and continue to dictate approaches used to develop HAR algorithms. Algorithms developed based on still images have been extended in video-based action recognition. These algorithms for visual-based action recognition originate from computer vision which is a field of artificial intelligence that programs computers to interpret and understand the visual world as human beings do.

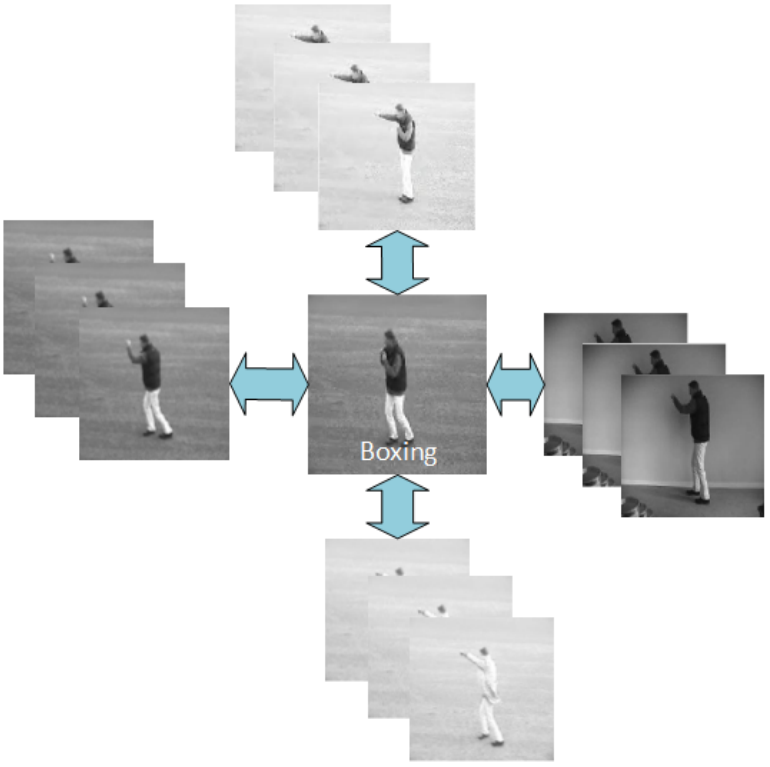


FIGURE 1.1: Highlighted challenges in action recognition: variant clothes, variant speed, variant background and variant illumination.

As we equip computers with algorithms to recognize human actions to overcome these challenges, one need to consider an aptitude, proper to human beings, which is not just related to the assumed knowledge, but also to the logical reasoning and the ability to extract relevant information as feature descriptors from a given context. In doing so, many methods proposed in the past two decades can be categorized into two main approaches: the handcrafted feature based approach and the deep learned feature based

approach. Handcrafted feature based techniques have been the most commonly used in computer vision. They are also known as traditional approaches and require knowledge of the data to inform the extraction of relevant features. Various methods such as Scale-Invariant Feature Transform (SIFT) [31], Histogram of Oriented Gradients (HOG) [32], Local Binary Pattern (LBP) [33] and Slope Pattern Spectra (SPS) [34] have been proposed to extract spatial or spatial-temporal features of actions. Though the handcrafted techniques in some cases have achieved high performance in action recognition, they remain highly problem-dependent and constrained to real-world applications. Contrary to traditional approaches, deep learning (DL) techniques also known as modern approaches because of their recency, build high level of representation directly from the data or raw video input without any pre-processing step as it is in the case for handcrafted feature-based techniques. These methods are based on the deep neural networks and rely on the adaptation of multilayered neural deep architectures to process real-world data [14]. The Convolutional Neural Network (CNN) is a popular DL model that comes in various dimensions depending on the data at hand. The 2D-CNN models handle 2D raw input frames which, in a HAR exercise, are referred to as videos frames. Such 2D-CNN models only learn spatial information while ignoring the motion information encoded in multiple adjacent video frames. Furthermore, 3D-CNN models are used to incorporate the motion information by capturing both spatial (2D) and temporal (1D) dimensions composing videos. However, 3D-CNN models are computational costly, giving room for simplified deep neural network models for HAR, especially in absence of adequate computational resources such as fast central processor units (CPUs) or graphical processor units (GPUs). Hence there is a need of a simplified DL model with comparable recognition performance that addresses the problem of lack of motion information when applying the 2D-CNN model and the computational cost of 3D-CNN model due to the denseness of the DL network.

Both handcrafted and deep learned methods produce feature representations which can be either local or holistic. Each of these attributes has an impact of the HAR performance. Methods based on local representation encode a sequence of video as a collection of local spatio-temporal features also described as local descriptors. These are extracted from spatio-temporal interest points (STIPs) which are sparsely detected from video sequences [35][36][37]. Human body parts in the case of video and regions of interest (ROI) in the case of still image are targeted and tracked while extracting the local features. This exercise in practice is computationally expensive. Although these local feature descriptors are discriminative enough and in some applications yielding good recognition accuracy, they lack the ability to capture adequate spatial and temporal information contrarily to the holistic representation which treats a still image or a video sequence as a whole. Holistic representations with their ability to encode visual information while

preserving both spatial and temporal structures of actions in a video sequence will be our main interest. Three types of holistic feature-based representations have been identified in the literature; statistical feature-based representation, structural feature-based representation and deep learned feature-based representation, and will be discussed in more detail in this thesis.

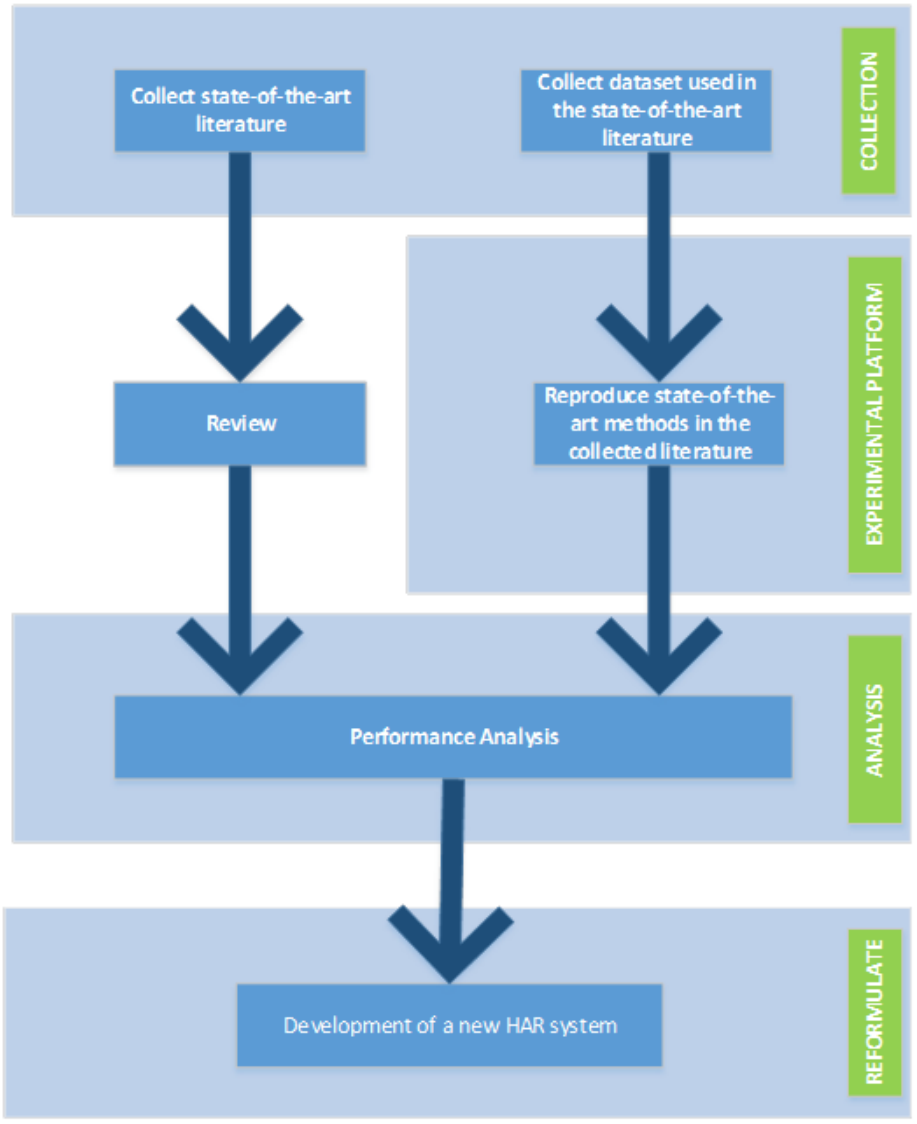


FIGURE 1.2: Research design.

1.2 Problem statement

Handcrafted techniques, by their hand-engineered features, often provide a right compromise between action recognition accuracy and computational efficiency. On the other side, the key success of DL techniques lies in their ability to learn better representation mostly from unlabeled data. As human movements are performed at different level of

abstraction, HAR can be referred to as the analysis of human motion from images or videos. Both handcrafted and DL techniques for HAR present a main challenge which is the finding of a proper representation of actions. By *proper* we refer to features of low dimension which are discriminative and able to generalize learning algorithms for the recognition or classification of all considered actions, including the unseen or untrained samples. In the effort to extract *proper* representation of actions in the literature, an increase of dimensionality of feature descriptors was produced without an increase of the number of training samples, leading to a phenomenon called the curse of dimensionality. The aim of this thesis is to develop new holistic approaches for action representation by adopting vision-based techniques to HAR ranging from handcrafted to DL- based techniques. This is achieved by creating features with reduced dimension which lessen the computational burden of learning models such as k-nearest neighbor (KNN), support vector machine (SVM) and softmax classifiers.

1.3 Research aim and objectives

This thesis focuses on the problem of action recognition in visual materials such as images and videos with main goal to explore state-of-the-art holistic approaches and propose effective HAR systems with very low computational complexity. In order to achieve our aim, the following specific objectives as illustrated in Figure 1.2, have to be met:

- Review the state-of-the-art approaches applied to HAR with more emphasis on holistic feature-based representations. Related limitations of the state-of-the-art techniques are understood and research gaps identified.
- Set up an experimental platform as depicted in Figure 1.3 that will allow us to reproduce the targeted state-of-the-art related articles in order to understand the methods and their respective limitations.

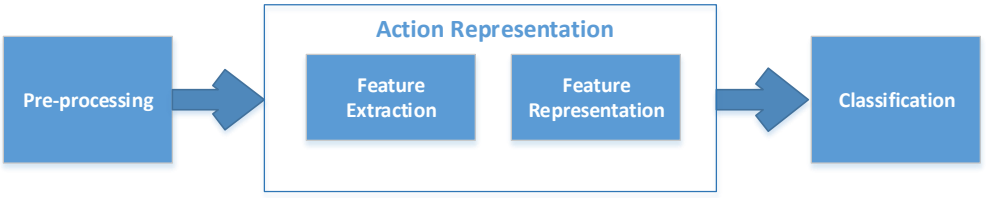


FIGURE 1.3: Experimental framework of a typical HAR system.

- Evaluate and compare the state-of-the-art methods in terms of both recognition rate and computational complexity.
- Develop new HAR systems that overcome the limitations of state-of-the-art methods.
- Evaluate and compare new developed methods to the state-of-the-art methods in terms of both recognition accuracy and computational complexity.

1.4 Motivation

Human action recognition is an important topic in computer vision due to its many related applications such as video surveillance, health care, and human-computer interaction. As visual technology advances, camera devices are found almost everywhere where a scene need interpretation. Most of the video surveillance or health care systems have been operating manually in the sense that they required a human expert to detect, track, recognize and analyze human activity events. Such monitoring systems, due to the absence of human experts in the control room, have yielded deadly consequences such the drowning of a child, falling down of a pedestrian in a traffic, neglect of elderly living alone, and occurrence of suspicious activities in public places such as banks, airport terminals, shopping centers [38–40]. Although an automated HAR system is beneficial to tackle such circumstances, its operating running time or time response in recognizing an action or interpreting a scene is even more important for a timely intervention. However most of HAR systems in the literature were developed by giving more attention to the recognition accuracy while ignoring the extensive computational cost. Since the computational complexity of HAR system is a great concern for real-time applications, and is associated to approaches used for action representation, holistic approaches are researched in this study because of their production of robust features at a lower computational cost. These approaches do not require the localization of the body parts but consider the whole body structure and dynamics to represent human actions. They are also referred to as appearance-based approaches and used in combination with other features extraction techniques to recognition human actions.

1.5 Main contributions

In this thesis, after exploration of action representation, algorithms related to feature extraction and representation of actions are proposed for their importance in finding

an effective and computational efficient human action recognition. Feature engineering played an important role. First, our proposed handcrafted algorithms aimed at characterizing intrinsic spatial-temporal features of human actions in a way to achieve comparable state-of-the-art recognition accuracy at low computational cost. Second, a pre-processing step is proposed to DL technique to reduce the computational complexity. Hence we briefly present our research contributions as follows:

1.5.1 Handcrafted-based contributions

- *Motion History Image - Slope Pattern Spectra (MHI-SPS) method*: This method is proposed after observation of the presence of slope patterns in MHI images describing the human movements in videos. The SPS feature extraction technique proposed in [43] easily capture slope information in MHI images and produce very low dimensional feature vectors which are later fed into a KNN classifier in order to recognize various human actions. Hence the combination of MHI represention and SPS feature extraction performed better than the solely use of MHI for HAR [34].

- *Circular Derivative Local Binary Pattern (CD-LBP)*: As a novel feature descriptor, CD-LBP is able to capture the appearance information efficiently. This proposed algorithm takes advantage of the circular binary structure of the LBP texture operator introduced by Ojala et al. [41] to evaluate the relation between consecutive binary digits. The CD-LBP approach provides higher-order CD-LBPs as lower dimensional feature descriptors which yield a flexibility in deciding on a good trade-off between recognition accuracy and running time. Good results were achieved on FER application [42].

- *Motion History Image - Circular Derivative Local Binary Pattern - Histogram of Oriented Gradients (MHI-CDLBP-HOG) method*: The CD-LBP feature extraction technique is also applied to MHI images to address the issues of self-occlusion. MHI templates are translated into CD-LBP codes which are at their turn translated into HOG features. The HOG features are later classified using the SVM classifier to produce state-of-the-art results where considering the Pedestrian action dataset. Although this method is computationally costly because of the involvement of two feature extraction techniques, it is still better with CD-LBP codes than LBP codes.

1.5.2 Learned-based contribution

Deep neural networks possess the ability to operate on raw video inputs. A 3D-CNN approach was best suited for HAR applications in the past. Here we propose a 2D-CNN for HAR application which take advantage of the MHI representation to learn robust features from temporal information embedded into the motion history images of videos. In doing so, the computational complexity imposed by the 3D-CNN approached is then reduced in the proposed 2D-CNN approach. Good results are found and compared even favorably against state-of-the-art handcrafted approaches.

1.6 Thesis outline

The rest of this thesis is structured as follows:

Chapter 2 provides relevant literature review. A brief introduction to local action recognition is presented for completeness purpose on the topic of human action recognition. Holistic approaches along side with their respective application and limitations are discussed. As feature extraction and representation constitute action representation, their distinction is highlighted. Learned models such as KNN, SVM and softmax classifier are also reviewed. Datasets used in the literature are also presented.

In Chapter 3, a novel feature descriptor named CD-LBP and two CD-LBP-based methods are proposed to describe appearance-based patterns. This new feature extraction technique is first applied on still images to describe the human emotions such as happy, sad, disgust, angry, fear, and neutral . This approach is then extended to HAR application where it is used in combination with MHI templates and other feature descriptors such as SPS and HOG. JAFFE dataset is used to validate our proposed technique when applied to still images, and both the KTH and Pedestrian action dataset when applied to videos.

In Chapter 4, we present DL approaches. The 3D-CNN based approaches are discussed in details. In addition, a pre-processing step introduced in the CNN scheme yielding a 2D-CNN approach for HAR is presented. For validation purpose, the KTH dataset is used to validate the proposed 2D-CNN approach and results provided.

Lastly, conclusion and future work are highlighted in Chapter 5. The overall findings of this thesis are summarized and drawn together, followed by a discussion that leads to an outlook for future research directions.

Chapter 2

Literature review

"All knowledge is connected to all other knowledge. The fun is in making the connections."

Arthur C. Aufderheide

2.1 Introduction

Human action recognition has now become a very interesting topic in the field of computer vision. The term *action* is often used in an interchangeable manner with the term *activity* by many authors in the literature. In this thesis, by action we refer to as a single person performing, a simple motion pattern executed by a person that lasts for a very short time duration [44], or a change brought by an action to the environment [45]. Although an activity also refers to as a sequence of actions [46], we only focus here on a single action performed by a person and captured by a camera. However, other HAR approaches have used non-visual sensors or wearable sensors to be specific, but presented limitations such as the discomfort felt by the user from wearing the sensors, reduced battery life, lack of context and unsuitability in video surveillance. Therefore the adoption of visual-based HAR approach is appropriate and constitutes our rationale. This type of approach, which has been used in a wide range of applications, is counted among the most popular HAR approach in the computer vision research community. The last decade have got researchers very occupied on HAR or related applications from both images and videos. Many surveys [1, 47–51] conducted on this field organize proposed HAR approaches into two categories: local approaches and holistic approaches. The proposed research topic will focus on the latter which is the holistic approach for the human activity recognition. This because the holistic approach produces more

robust features with better interpretability. Moreover, the state-o-the-art HAR techniques have shown good recognition rate but with an extensive computational cost. The computational complexity, though a great concern for real-time applications, has been overlooked in the recent state-of-the-art methods. Since such complexity is originated from the high dimensional datasets to be processed, we will then explore, with intention to identify potential gaps and develop new HAR systems, the feature extraction or dimensional reduction techniques that could lead to subspaces with both high discriminative features and less computational burden. Hence this chapter briefly discusses the notion of spatio-temporal features and local action recognition in Sections 2.2 and 2.3 respectively, and further provides a comprehensive review, analysis, evaluations, detailed comparisons, and important discussions on the most recent and prominent holistic features based techniques which are handcrafted feature-based methods in Section 2.4.1 and deep learning feature-based methods in Section 2.4.2. The classification techniques, which are widely used in action recognition, are discussed in Section 2.5. Publicly available datasets used in HAR systems are described in Section 2.6. Finally, the chapter is summarized in Section 2.7. The taxonomy for HAR presented in Figure 2.1 resumes the literature review.

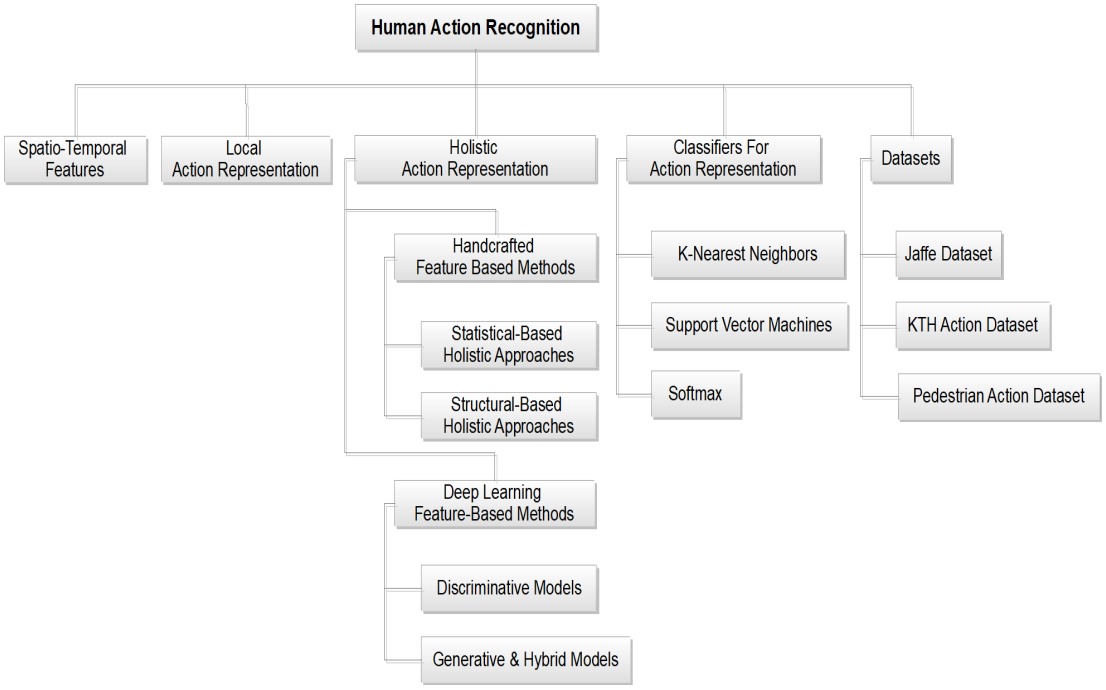


FIGURE 2.1: Taxonomy for Human Action Recognition

2.2 Spatio-temporal features in action recognition

Images are 2D data formulated from a projection of 3D real-world scenes and contain spatial structures such as shapes and appearances of humans and objects. These images, placed in a chronological order, form a sequence called a video. So an executed action in a video sequence can be represented as a unique 3D space-time volume constructed by a concatenation of 2D images along the time axis. Such representation has not been beneficial in its raw data state for HAR application because of the large number of pixels the volume contains, which is the 2D image size times the number of frames considered on the time axis, and also the redundant information existing between consecutive frames. Therefore more compact action representations resulting to new and discriminative spatio-temporal features extracted from 3D space-time volumes are a requirement to tackle challenges such as variable illumination, background changes, viewpoint changes, and partial occlusions. Various human action representations have been extensively studied in the past two decades. Related action representation-based approaches are based on spatio-temporal information to distinguish between action observed from video sequences. These approaches can be categorized into two mainstreams. (1) Local action representation also known as local technique encodes a video sequence as an accumulation of local spatio-temporal features also referred to as local descriptors. This type of technique involves two subsequent steps which are feature detection and feature description. More about local action representation-based approach is briefly discussed in Section 2.3 for completeness purpose. (2) Holistic action representation also known as global technique processes a video sequence as a whole rather than dealing with detected local video patches. This type of technique has been solicited due to its ability to encode visual information while preserving both the spatial and temporal structures of the action occurring in a video sequence. Moreover, this technique often require a pre-processing step such as background subtraction and segmentation. However, previous spatio-temporal feature-based approaches whether local or holistic have focused on action recognition accuracy, and given less attention on the computational cost caused by these action representations. Local action representation is discussed in the next section.

2.3 Local action representation

The action representation discussed in this section are done through local approaches, also known as local methods, that use local spatio-temporal features extracted from 3D space-time volumes in order to represent and recognize actions. These approaches are motivated by the fact that a 3D space-time volume is essentially considered as a solid 3D

object and the extracted appropriate features characterizing 3D volumes for each action can be used to recognize an action through learning models. Generally, the recognition of an action is described by a methodology which engages three components of which two, feature extraction and feature representation, are used for action representation, and the third one is the classification component which has the ability to learn from the newly represented features. Moreover, local features are first extracted to capture local motion information of an human from the 3D space-time volume. Secondly, the extracted features, which may come on various forms, are then put together to represent the actions while taking into account or ignoring possible exiting spatio-temporal relations. Lastly, learning or recognition algorithms are used to classify the actions.

Furthermore, as local representations describe action as a collection of independent patches, their computation is done by detecting spatio-temporal interest points (STIP) followed by the description of local patches around these points [1]. Thus the extraction of local spatio-temporal features consists of two processes– feature detection and feature description. The former analyzes every pixels to detect the presence of a feature at the considered pixel. Around the detected pixels designated as STIP, patches are cropped at pre-determined spatial and temporal scales and orientations. This process is done with the help of feature detectors such as Cuboid [37], Harris3D [31], Dense sampling [52], and Hessian [53]. Figure .2.2 shows examples of cuboids at detected interest points. The

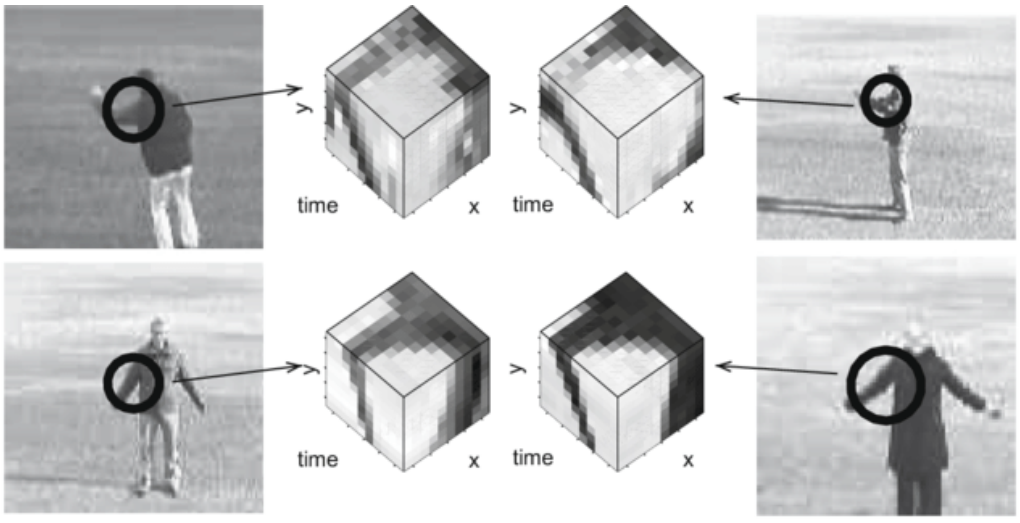


FIGURE 2.2: Extraction of spacetime cuboids at interest points from similar actions performed by different persons [1].

latter describes local patches in local representations that are ideally invariant to background clutter, appearance and occlusions, and possibly to rotation and scale [1]. Few feature descriptors have been identified in the literature. In [37], the Cuboid descriptor is used to describe the detected 3D patches, and the gradients derived from them are concatenated into a feature vector which is later reduced by applying a dimensional

reduction such as PCA. In [31], a combination of HOG and HOF descriptors is used to describe the local appearance and motion by binning histograms of spatial gradient and optical flow which are computed and accumulated in space-time neighborhoods of detected interest points. In [2] and [32], SIFT and HOG descriptors, originally proposed for the recognition of objects and pedestrians in static images by Lowe [54] and Dala et al. [55] respectively, are extended to space-time volumes to extract dynamic features for the recognition of human actions in video sequences. These extensions also named 3D-SIFT and HOG3D descriptors share in common some similarities as they are both based on histograms of gradients. The process of extracting a HOG3D descriptor is shown in Figure 2.3. Here regular polyhedrons are used in a uniform way to quantize the orientation of spatio-temporal gradients. Another descriptor used for the extraction of spatio-temporal features is the extended SURF descriptor [53] which is simply an extension of the SURF descriptor used for 2D image but applied to videos. This

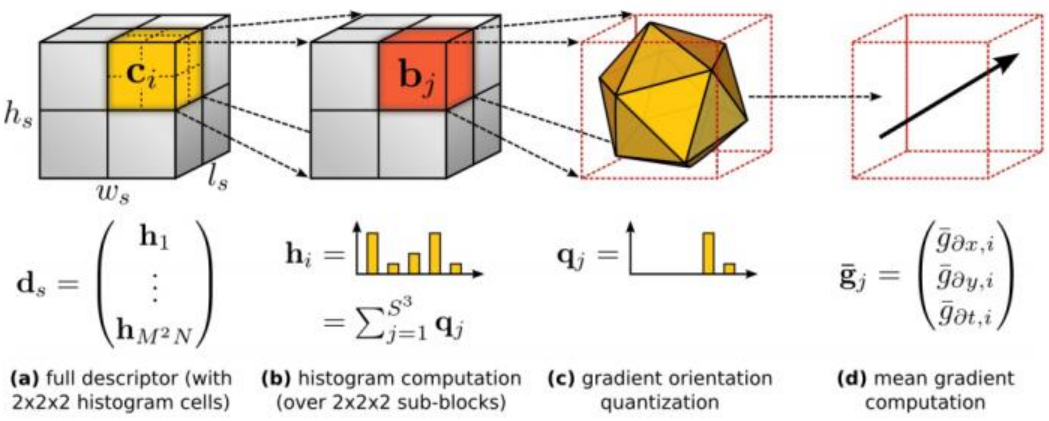


FIGURE 2.3: Process of extracting HOG3D descriptor [2].

descriptor also shares some similarities with the SIFT descriptor but differs to other descriptors with its quantization measure leading to the feature vectors. In addition, new spatio-temporal features resulting from the above-mentioned local descriptors, because of their sparse or dense nature and for classification purpose, are further processed into a suitable dictionary of visual vocabularies through feature representations such as k-means, Bag-of-words (BoWs) and sparse coding. Besides propagating large quantization errors, the computational cost of feature representations is as expensive as the process of detecting and describing local features. For that reason, we also review in the next section the holistic action representation for its good interpretability, simplicity and comptutational efficiency.

2.4 Holistic action representation

In this section, we discuss various holistic action representations also known as holistic or global approaches. A HAR system developed based on a holistic approach is referred to as a recognition system that takes into account the whole videos sequences and extract holistic features that are later used in the classification. The term *holistic* implies the homogeneity in representing the spatio-temporal descriptors. Here every point in the space-time volumes is processed in the same way. This consequently eases up computation of features which is done densely or parallelly. Recently, besides their computational efficiency, these representations have drawn an increasing attention due to their ability at encoding more visual information while preserving both spatial and temporal structures of occurring actions in video sequences. In the literature, these representations can be categorized as two type of methods, the handcrafted feature based methods and the deep learning feature based methods.

2.4.1 Handcrafted feature-based methods

Handcrafted feature-based approaches also known as traditional approaches have been the most commonly used in computer vision. These approaches require a knowledge of the data at hand in order to extract relevant features; features that are discriminative and of reduced dimension for good recognition. Primarily, they were applied to 2D data such as images and then later extended to 3D data to accommodate vision-based human action recognition. Such extension has facilitated the search for interest points known as STIPs [31] in videos leading to local action representations as discussed in 2.2, and the exploitation of the whole 3D space-time volume leading to holistic action representations [56] known also as global approaches. In contrast to local action representations which are based on the description of human actions as a collection of independent video patches, holistic action representations are referred to as methods based on the global appearance and motion which capture global features irrespective of any notion of parts of the human body or STIPs. As handcrafted feature-based methods generally operate at video pixel levels and measure low-level statistics of spatial body shapes and temporal motions [51], handcrafted holistic approaches in the literature can be grouped into statistical feature based approaches and structural feature based approaches.

2.4.1.1 Statistical-based holistic approaches

Statistical-based holistic approaches such as PCA, LDA, FLDA and RDA have tremendously been employed in the field of pattern recognition, especially in face recognition

(FR) [57–62]. Most of these techniques were copied from FR and improved to suit the recognition of human activities. These techniques, also identified as feature extraction, play an important role in reducing the high dimensionality of the data.

The PCA approach have been proven successful in the FR related applications. This approach transforms a two-dimensional facial image expressed in one-dimension into compact principal components representing the feature space [63]. These principal components are essential characteristics known as eigenfaces [64]. The derived features are arranged with respect to their respectively importance. The least important features belong to the null subspace of the eigenspace. By removing features partly (null subspace), the approach acts as a dimensionality reduction technique. However the features obtained from this approach lacks robustness and yields to poor recognition due to inadequate feature representation. When applying PCA to HAR, the initial data (video) is then mapped into a lower dimensional space through a linear combination of the best eigenvectors. These eigenvectors are derived by retaining eigenvectors corresponding to the highest eigenvalues from the computed sample covariance matrix. The computation of covariances is effectuated from the considered dataset. Let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of $m-$dimensional column vectors constituting the training set taken from the dataset , where $x_i \in R^n$. The covariance matrix $S \in R^{m \times m}$ can then be described as

$$S = \frac{1}{n} \sum_{i=1}^{n} (x_i - c)(x_i - c)^T, \tag{2.1}$$

where $c = \frac{1}{n} \sum_{i=1}^{n} x_i$ is the average of all the data points. As the matrix S is real and symmetric, then there exist a real and unitary matrix V and a real diagonal matrix Λ such that $V^T S V = \Lambda$. The elements of the diagonal matrix Λ are the eigenvalues of S, and the columns of V are corresponding eigenvectors of S also known as principal components. Considering $\lambda_1, \lambda_2, \ldots, \lambda_n$ sorted as $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_n$ and their respective eigenvectors $v_1, v_2, \ldots, v_n$, a new feature vector or data point is formed by projecting the original data point x_i on the eigenvector v_k by means of

$$y_{ik} = v_k^T (x_i - c), \quad 1 \leq k \leq n. \tag{2.2}$$

Moreover, a collection of feature vectors y_{ik} sorted in decreasing order of eigenvalues, represents a new coordinate system where projected data points are well observed. Based on the fact that eigenvalues near zero are perceived to hold less relevant information for recognition purpose, their corresponding vectors are discarded [65]. Hence a reduced coordinate system constituted of eigenvectors with the highest eigenvalues is created.

The considered set of eigenvectors is denoted by

$$V = [v_1, v_2, \ldots, v_l] \tag{2.3}$$

with $l < n$ and l being the number of features often selected based on an application.

In [3], PCA was used as an unsupervised approach to extract features from the Weizmann database which were later classified using the Euclidean distance measure but performed poorly. Contrarily to the direct application of PCA as a feature extraction techniques, PCA is the most used data reduction technique in the literature. It has been used either in the preprocessing or the post-processing stage of HAR systems. In [66], the pre-processing of silhouette vectors was followed by the PCA technique which extracted the relevant human activity silhouette features by decreasing the amount of redundant information. Other feature extraction techniques used in association with PCA for HAR are HOG, Local Binary Pattern on Three Orthogonal Planes(LBP-TOP) and 3D-SIFT which were found in [67], [68] and [69] respectively. Though the use of PCA in a post-processing mode yields very reduced feature samples for the learning models, it increases the computational cost of the overall HAR system.

Another statistical based holistic approach that betters the recognition rate produced by PCA is the LDA approach. This approach also known as a generalization of FLDA is a supervised technique that outperforms the PCA approach in terms of classification performance by taking class separability into consideration [70]. Linear discriminant analysis determines, while reducing the dimensionality of feature vectors, a projective feature space with most discriminant features. This is achieved by maximizing the between-class scatter matrix and minimizing the within-class scatter matrix. These matrices are respectively denoted by S_ω and S_b, and respectively defined by

$$S_\omega = \sum_{i=1}^{nC} \sum_{j=1}^{nE} (y_j - c_i)(y_j - c_i)^T, \tag{2.4}$$

and

$$S_b = \sum_{i=1}^{nC} (c_i - c)(c_i - c)^T, \tag{2.5}$$

where nC, nE, y_j, c_i and c represent respectively the number of classes, the number of elements per class, the resulting feature vectors derived through PCA by means of Equation 2.3, the mean of each class, the mean of all the classes. The optimal discrimination matrix W_{lda} that maximizes the ratio of the determinant of the between-class matrix S_b to the determinant of the within-class matrix S_ω can be found by solving

the following optimization problem

$$W_{lda} = \arg max \frac{W^T S_b W}{W^T S_\omega W}, \tag{2.6}$$

where W_{lda} is the set of discriminant vectors of S_w and S_b corresponding to the (nC-1) largest eigenvalues. The objective function in Equation 2.6 implies the need to increase the between-class variance and decrease the within-class variance. Thus, if S_w is not a singular matrix, then W_{lda} can be found by solving the equivalent generalized eigenvalue problem defined by

$$S_b W_{lda} = \lambda S_w W_{lda}, \tag{2.7}$$

which implies that

$$S_w^{-1} S_b W_{lda} = \lambda W_{lda}. \tag{2.8}$$

Moreover, the set of discriminant vectors denoted by $W_{lda} = [w_1, w_2, \ldots, w_k]^T$ can be found by solving Equation 2.9 below

$$S_b w_i = \lambda_i S_w w_i, \quad i = 1, 2, \ldots, nC - 1 \tag{2.9}$$

where the rank of S_b is less or equal to $nC-1$ with $nC-1$ which is the maximum value of k. Hence the feature vectors produced by the LDA approach are given by multiplying the new feature vectors derived from the PCA (Equation 2.2), which is an intermediate step of LDA, by the discriminant vectors (2.9). This technique has produced very good classification results in a number of applications [71–73], but still suffers from two major problems (1) small sample size (SSS) problem which is experienced when the dimension of the reduced feature space is still larger than the total number of the training vectors, and (2) the null feature space which may possess discriminative properties but ignored during the feature representation. In [3], it was found that the result of the discriminant evaluation according to Fisher criteria was not directly practical to the HAR area due to the limited number of training samples, temporal dependencies in the image capture of the same activity, and the noise in high dimensionality of the activity images. Another implication to this impracticability was the computation of the inverse of the within-class scatter matrix as required in Equation 2.8. An inaccurate computed result produces a imprecise estimation of the eigenvalues leading to an exponential increase of error distorting the discriminant analysis. High computational complexity was also observed in many solution discussed in [70]. Among these proposed solutions, RDA also referred to as a generalization of LDA, has been proven to be more effective [74–76], and robust against the SSS problem [77]. Initially introduced by Friedman in [78], the first RDA approach is based on the shrinkage covariance matrix with two parameters found through a grid-searching process minimizing the classification error. Later on, there have been

many progressive works done which produced RDA variants with good improvement of recognition performance despite high computational complexity.

Recently, related RDA techniques, which involve both eigenspectrum modeling and subspace decomposition, have been applied to recognize human activities. These techniques aim at decomposing the entire eigenspace spanned by the eigenvectors into subspaces. An illustration of a decomposition into three subspaces is depicted in Figure 2.4. These subspaces are a reliable activity variation-dominating subspace, an unreliable noise variation-dominating subspace and a null subspace, and denoted respectively by $P = \{\varphi_k^w\}_{k=1}^{m}$, $N = \{\varphi_k^w\}_{k=m+1}^{r_w}$, and $\Phi = \{\varphi_k^w\}_{k=r_w+1}^{n}$. Moreover, the decomposition of the eigenspace allows the identification of unreliable eigenvalues and their modification for better generalization through parametric eigenspectrum modeling. In [3, 79], the authors have proposed a 3-parameter based eigenfeature regularization scheme (3-parameter) for HAR application which was an improvement of the 2-parameter applied to FR [59]. This 3-parameter scheme outperformed both the 2-parameter and the FLDA when applied to popular database such as KTH, Wizmann datasets [3]. Furthermore, in [80] a proposed 4-parameter regularization scheme that outperformed the 3-parameter on the KTH databases in terms of recognition rate, the calculation of the parameters added a computational load to the whole HAR process. However, the computational

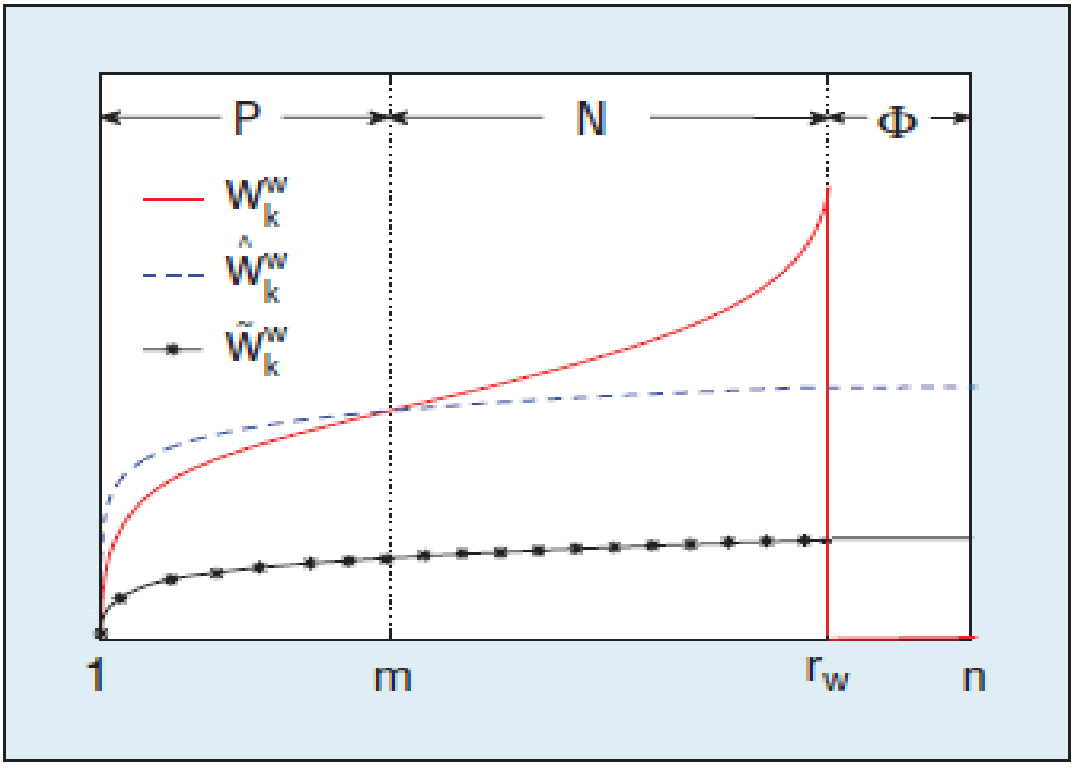


FIGURE 2.4: Eigenspace decomposed into principal (P), noise (N), and null (Φ) subspaces based on a typical real eigenspectrum of human activity data [3].

complexity of the proposed RDA methods for HAR application is greater than other

subspace methods because the entire space of the within-class scatter matrix is considered while evaluating discriminant values. By consequent, statistical-based approaches are computational costly, especially the search of best-fit parameters. Hence the need to explore other holistic or appearance-based approaches.

2.4.1.2 Structural-based holistic approaches

Structural-based holistic approaches are techniques that represent human actions by their global appearance and motion while preserving their structural information made of both spatial and temporal structures of actions taking place in a video sequence. Action representations produced from these approaches have shown satisfactory performance, especially in scenes with relatively static background [7, 81, 82]. Most of HAR methodologies employing holistic approaches require a pre-processing stage to get visual data gathered from visual-based sensors ready for further processing as shown in Figure 1.3. This stage includes image processing tools such as background subtraction [83], foreground extraction [84], and morphological operations [85]. This stage is also crucial in holistic HAR applications as it allows isolation of the moving subject in the video by reducing the research space which results in a region of interest (ROI), and improve performance in terms of computational cost [86]. Figure 2.5 shows an example of moving human detection using background subtraction technique which is the simplest of all. This technique consists in separating the foreground from the background image in

FIGURE 2.5: Example of background extraction

order to get an image that contains only the moving object. A subtraction with respect to a predefined threshold is done between every new image and the initial background image to obtain relevant information referred to as silhouette information. However, a static background image contributing to an effective extraction of the background does

not always reflect the reality since real-life applications in most cases experience a dynamic background which are results of change in illumination, moving objects that are not part of the ROI, presence of shadows, and also moving viewpoints of the camera. Many methods on background modeling have been proposed to address the problem of robustness [87–90]. In addition, the detection of ROI has given rise to the creation of three types of global approaches to represent human action holistically. The first one is based on the human shape derived from the pre-processing stage. This approach is also called *silhouette-based approach* since it targets the 2D spatial representation of actions. The second one is based on the human motion displayed by consecutive silhouettes. This approach is also known as *motion-based approach* since it targets the motion structure of actions. The third one is based on the spatio-temporal volumes formed by stacking silhouettes. This approach is then referred to as *spatio-temporal-based approach* since it targets the combination of both the spatial and motion structures.

Silhouette-based approaches characterize actions in video by capturing shape features from the human image or silhouette obtained from background extraction. Although a human silhouette is subject to noise due to an fallible background extraction and somewhat sensitive to different angles at which video is captured, it still encodes implicitly the anthropometry of the human while carrying important information such as contour and the occupied region. In [91], relevant features are extracted from the silhouette and its surrounding regions between canvas and human body. The work of Yamato et al. [4] is among the first to propose silhouette images (see Figure. 2.6) for

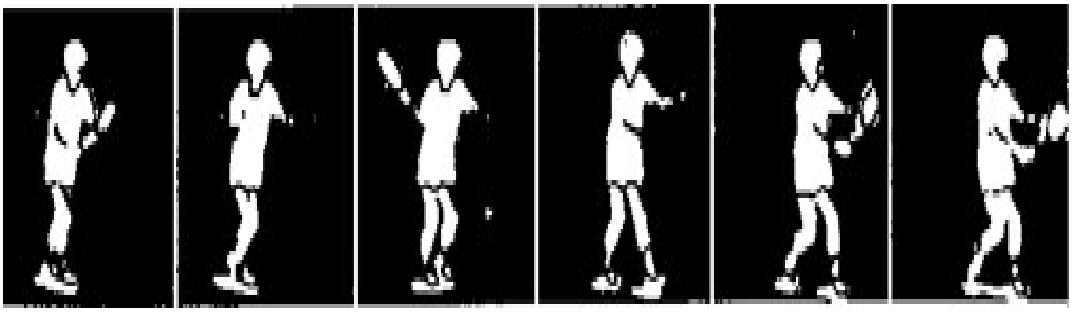


FIGURE 2.6: Examples of human silhouettes for tennis play actions [4]

HAR based on Hidden Markov model (HMM). A grid representation is computed over the silhouette and a ratio of foreground to background pixel for each cells within the underlying region of the silhouette is calculated. These grid representations are then quantized into a vocabulary, and human actions learned as sequences of words using a HMM. In [92], the authors have proposed matching template techniques that reduce significantly the effect of noise in silhouette image by dividing the ROI into fixed spatial or temporal grid. Moreover, a region-based HAR approach is proposed in [93], where the human silhouette is divided into fixed number of grids and cells for action representation

which is later fed to a hybrid SVM-NN classifier for the purpose of action recognition. Another approach proposed in [94] has exploited the contour points of the human silhouette in a radial scheme to represent action and the SVM classifier used for classification. These contour points also served for global pose representation in [95], where action were learned with sequences of key poses. Such pose representation helped to deal with issue of redundancy introduced by the inside part of the silhouette. Also no morphological pre-processing steps were needed to better the detection of the human silhouette and reduce the sensitivity due to variations in viewpoint or illumination. However, although the silhouette-based approaches provide shape information they fail to capture motion information which also play a key role in describing actions in videos.

Motion-based approaches form part of holistic or global approaches for HAR. These approaches characterize actions in a video by capturing motion features between successive video frames. They are implemented through either extraction of optical flows or accumulation of the human silhouettes. Approaches based on optical flows often represents motion in videos by measuring pixel displacements between two successive images. Here the motion information from the input video is transformed to feature vectors which are further fed into learning models. Polana and Nelson first employ optical flow method to track human along with an action representation using spatio-temporal grids of optical flow magnitudes [96]. In [5], a motion descriptor based on both smoothed and aggregated optical flow measurements over a spatio-temporal volume is proposed. The authors track a soccer players in a sport footage and calculate optical flow in the ROI. Derived descriptors, which are presented into four channels as shown in Figure 2.7, are used in a nearest-neighbor framework for classification purpose. Similar approaches for action representation are used in [97–99]. However, the success of using optical flow-based approach relies on the selected algorithms for a reliable optical flow estimation. Usually, features extracted immediately from the optical flow are imprecise since they are negatively influenced by environment variation such as illuminations, and noise [100]. For this reason, some improvement on the extraction of motion-based features are proposed in [101, 102]. Moreover, state-of-the-art performance have been achieved using optical flow to the expense of facing serious difficulties in estimating optical flow and resulted in an increased computational cost.

Spatio-temporal-based approaches originate from accumulated human silhouettes that result in compact forms called spatio-temporal motion templates or simply temporal templates. These approaches based on temporal templates provide global features as global action representation by focusing on human movement. The development of temporal templates such as Motion Energy Image (MEI) and Motion History Image (MHI) is originally introduced by Bobick and Davis [81, 103, 104] in order to recognize various types of aerobics exercise [81]. The MEI is a binary image that shows the regions

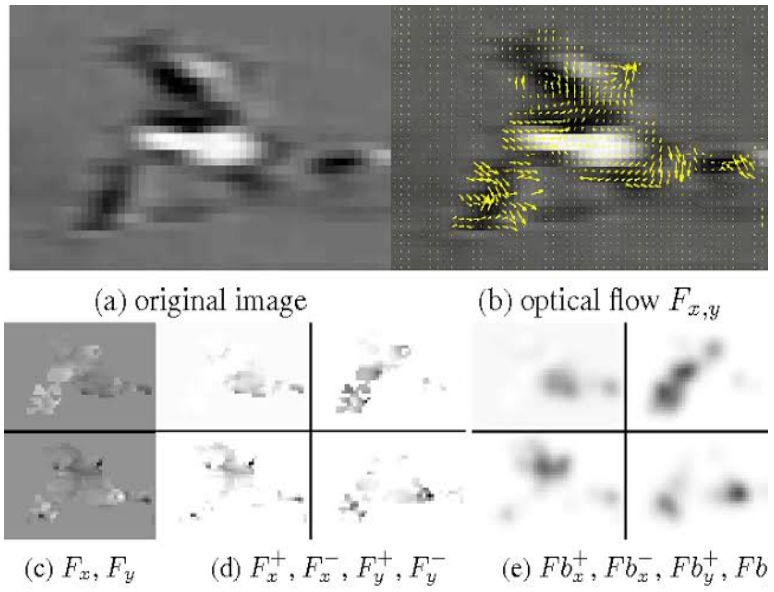


FIGURE 2.7: Constructed motion descriptor based on optical flow [5]. (a) Original video frame, (b) Optical flow $F_{x,y}$, (c) Separation into x and y components of optical flow vectors, (d) Rectification of channels producing 4 separate channels, (e) Blurred motion channels

where motion occurs, whereas the MHI is a grayscale image representing *how* motion changes over time. From Figure 2.8, one can observe that the MEI image present two

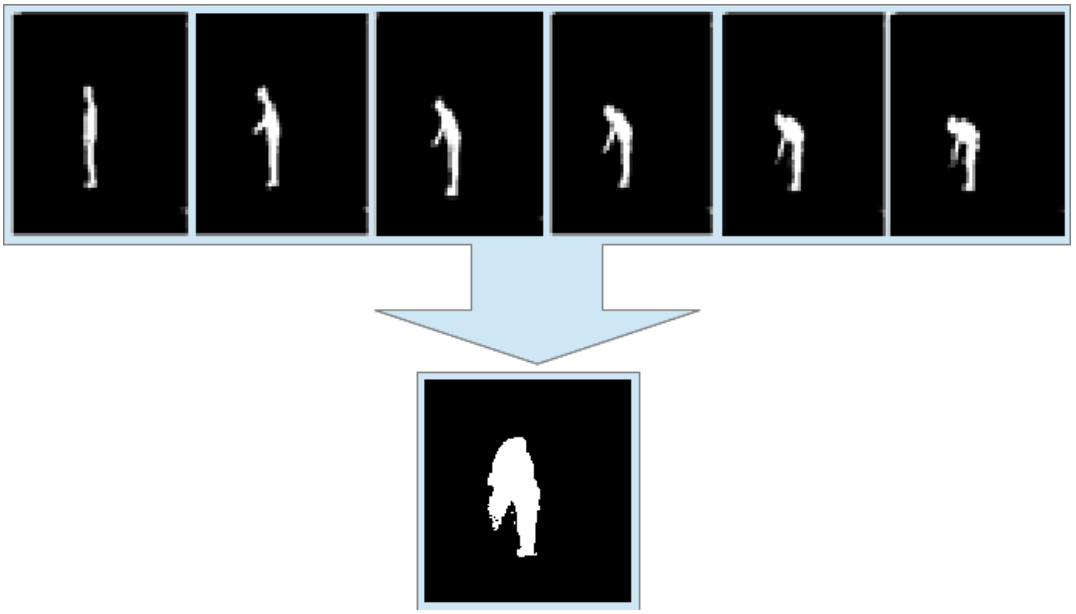


FIGURE 2.8: Example of a binary MEI image for a *bending* action

pixel intensity values, either '1' for white pixels indicating the motion regions referred to as energy of the image or '0' for black pixels indicating non-moving regions. Thus the MEI is then defined as spatial distribution of motion energy derived from a cumulative silhouettes. Let $I(x, y, t)$ be an image sequence and $D(x, y, t)$ a binary image sequence

representing motion regions which can be obtained by background subtraction or optical flow. Then the MEI denoted by $E_\tau(x, y, t)$ is defined by

$$E_\tau(x, y, t) = \bigcup_{i=0}^{\tau-1} D(x, y, t - i), \tag{2.10}$$

where the duration τ is critical in defining the temporal extent of an action [105]. However, the MEI representation provides the presence of motion but fails to show the direction of motion within its template as the MHI template does.

Moreover, based on the MEI template which shows the spatial distribution of the contours and energy of the motion, a Motion History Image (MHI) can be generated. The MHI template is a vision-based template representation that expresses the target motion in the form of image brightness by calculating the pixel change at the same position in the time period. It is an image in which the gray value of each pixel represents the most recent motion of the pixel at that location in a set of video sequences. The closer the last motion is to the current frame, the higher the gray value of the pixel. Therefore, the MHI image can represent the most recent movements of the human body during an action, which makes MHI widely used in the field of motion recognition. Figure 2.9 shows an example of MHI for a *bending* action. The concept of the MHI template is

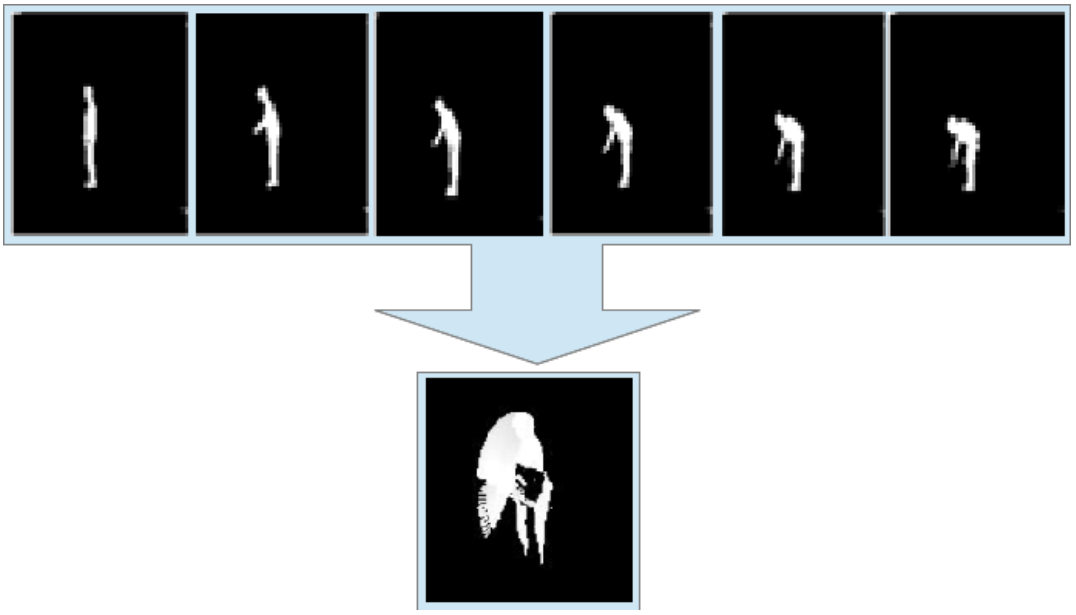


FIGURE 2.9: Example of a grayscale MHI image for a *bending* action

extensively discussed in [106, 107]. Let H_τ be the intensity value of the motion history

pixel, $H_\tau(x, y, t)$ calculated as

$$H_\tau(x,y,t) = \begin{cases} \tau & \text{if } \Psi(x,y,t) = 1 \\ max(0, H_\tau(x,y,t) - \delta) & otherwise \end{cases} \tag{2.11}$$

where (x, y) and t are the position and time of the pixel respectively; τ is the duration, which determines the time range of motion in terms of the number of frames; δ is the decay parameter. $\Psi(x, y, t)$ is an update function, which can be defined by various methods such as interframe difference, image difference or optical flow [108]. The interframe difference method is most commonly used and the update function defined as

$$\Psi(x,y,t) = \begin{cases} 1 & \text{if } D(x,y,t) \leq \epsilon \\ 0 & otherwise \end{cases} \tag{2.12}$$

and the interframe difference which is the difference of frames denoted by $D(x, y, t)$ is given by

$$D(x,y,t) = |I(x,y,t) - I(x,y,t \pm \lambda))|, \tag{2.13}$$

where $I(x, y, t)$ is the intensity value of the t^{th}-frame coordinate (x, y) pixel of the video image sequence, λ is the interframe distance, and ϵ is the artificially given difference threshold, along with the video scene. All these parameters have a great effect in the generation of MHI images, therefore their selection should be done carefully in order to get a good motion representation. A comprehensive analysis is conducted in [6] to evaluate the dependences on parameters τ, δ, ϵ, and λ while developing MHI images. Figures

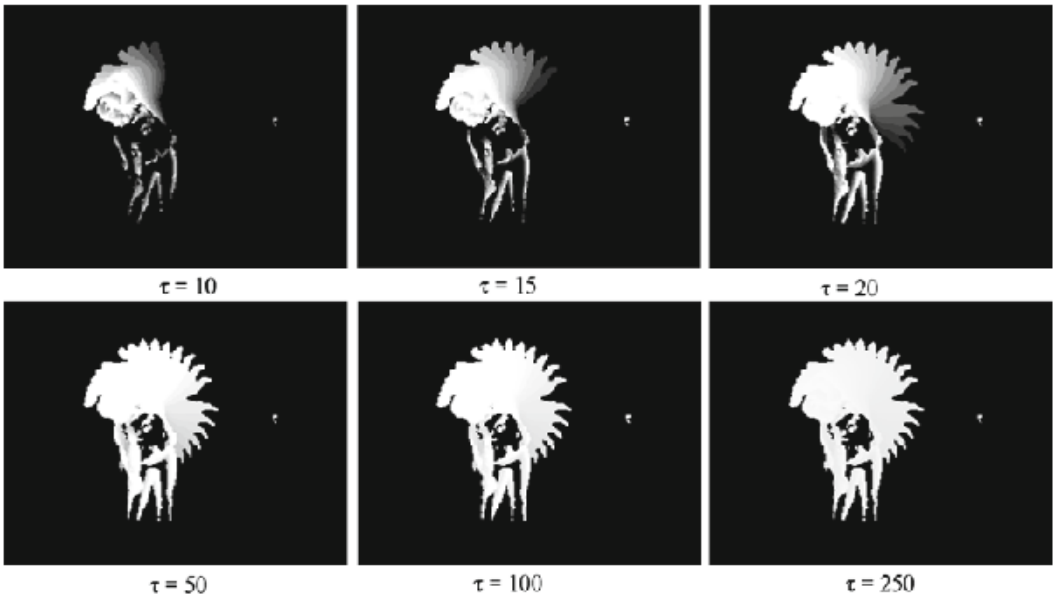


FIGURE 2.10: Illustraction of the dependence on τ in calculating the MHI template [6]

2.10, 2.11, 2.12 illustrate these dependences. When generating the MHI template, if the duration τ is less than the number of frames in which the motion continues, a loss of the motion information can be observed. As shown in Figure 2.10, for a *left-handed waving* action video made of 26 frames as length, the parameter τ equals to 10, 15, 20, 50, 100, and 250 respectively, and the parameter δ to be 1 are considered to make the MHI templates. For the case where the duration of motion as shown in the top row templates is less than the video frame length ($\tau < 26$) , the MHI template loses the motion information at the beginning of the motion. Conversely, if the duration τ is set too large compared to the video frame length, then the change in the pixel intensity value in the MHI template will become less significant. In the bottom row of Figure 2.10, the far right MHI template corresponding to a duration about 10 times the total length of the video ($\tau > 26$) shows the nonzero intensity values which are very close, making it difficult to distinguish between waving to the left and putting the hand down. Consequently, the impact caused by the selection of the duration τ with respect to δ must be considered when generating MHI images.

At the same time, the selection of the decay parameter δ has a significant impact on the generation of motion history templates. When reading the previous frame, for a particular pixel location where the motion region has occurred, if the location is turned to a stationary state or the motion state is not changed, the intensity value of the pixel in the motion history template is decreased by δ size. In the basic MHI representing method, the value of δ is usually taken as 1, but the actual δ value in the actual operation may change the information provided by the motion history image. Therefore, δ can take the corresponding empirical value according to the research needs. Figure 2.11, for

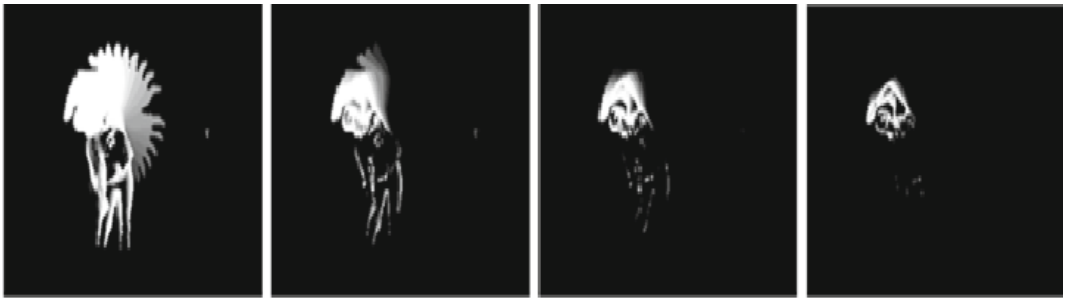


FIGURE 2.11: Illustraction of the dependence on δ in calculating the MHI template [6]

the motion of bending the left hand, illustrates the progressive removal of earlier trail of motion sequence as the δ value increases (1, 3, 5 and 10). In addition, the combination of the duration τ and the decay parameter δ in the motion history function determines the time at which the pixel intensity of the motion region decays to zero. A larger τ combined with a smaller δ produces a continuous, slowly varying gradient distribution, while a larger δ combined with a larger τ results in a discrete stepped stratification.

Moreover, the difference threshold ϵ is as well one of the important parameters of the motion history function. In the four MHI images shown in the Figure 2.12, the difference threshold increases from left to right, ϵ equals to 30, 50, 75, and 150 respectively. It

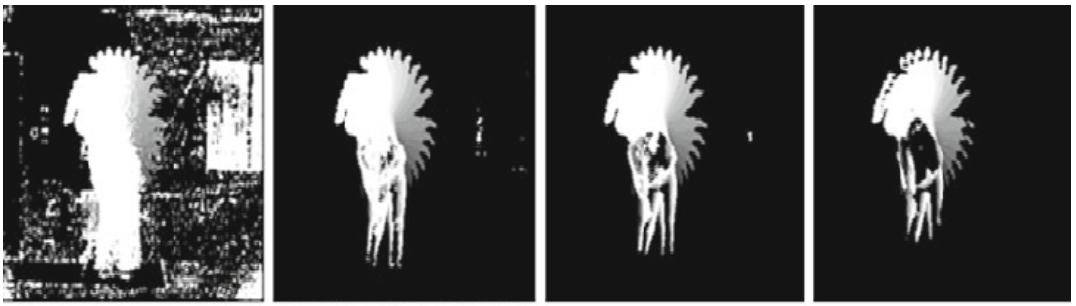


FIGURE 2.12: Dependence on ϵ. From left to right $\epsilon = 30, 50, 75, 150$ respectively [6]

can be observed that when the difference threshold is too small, the foreground and background of the motion cannot be well distinguished, and the background of the MHI template is full of noise. The background noise gradually disappears as the difference threshold increases, and instead a *cavity* appears in the center of the motion region. The void increases as the threshold increases until only the edge portion of the motion remains in the MHI template. However, for larger value of ϵ, some part of motion information is missed in the process. This is illustrated in the right-most image in Figure 2.12 for $\epsilon = 150$. Therefore, special attention needs to be paid on the selection of the difference threshold or the update function when handcrafting MHI templates.

Instead of modeling the motion history in a single image that is also challenged by issues of self-occlusions, Blank et al. [109] and Gorelick et al. [7] overcome this limitation by first stacking the silhouettes from a given video sequence to form a space-time volume also referred to as spatio-temporal volumes. Examples of spatio-temporal volumes for three different actions are shown in Figure 2.13. This approach, spatio-temporal

FIGURE 2.13: Examples of STV from stacked silhouettes [7]

volume-based approach, provides quite robust representation and utilizes properties of the solution of the Poisson equation to extract features such as local spatio-temporal saliency, shape structure, orientation of a pixel in relation to its neighborhood, and action dynamics. Global descriptors are thus obtained for a given time interval by adding

up the moments for local descriptors. Therefore the KNN classifier, fed by these features, facilitates the recognition of actions. Moreover, in [110, 111], silhouettes from multiple cameras are combined to build a representation based on voxels (3D pixels) resulting to a Motion History Volume (MHV) which is considered as an extension of MHI in 3D. Examples of motion history volumes for three different actions are shown in Figure 2.14. Although this representation is quite informative, it requires the extraction of silhou-

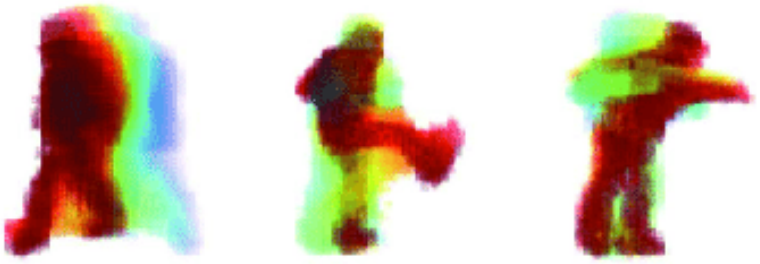


FIGURE 2.14: Examples of MHV from combined silhouettes [7]

ette and also a network of calibrated cameras. In [112], the authors have proposed to transform MHV into frequency domain. Prior to this, they first started by building the MHV which is divided into sub-volumes. The vector of primitives corresponds to the average frequency obtained. The classification is based on a matching technique using a weighted Euclidean distance. Their method has shown itself robust to both change of viewpoint and noise. However, the required processes such as parallel background subtraction, multiple camera synchronization, and mostly additional computational expenses due to calibration for STV approaches remain challenging [113] besides their satisfactory state-of-the-art performance.

Further, spatio-temporal templates (especially MHI templates) have been extensively employed to derive feature vectors as outcome from action representation for recognition purpose. Human action recognition forms part of applications where MHI-based methods are considered. The work of Ahad [106] demonstrates the employability of MEI, MHI and related variants in applications such as action recognition [6, 8, 81, 110, 114–122], gesture recognition [123–125], gait recognition [126–128] and video analysis [129–131]. Various strategies resulting to MHI variants are proposed for better motion representation, e.g. the Directional MHI (DMHI) tackles the issue of overwriting [119] and the Hierarchical Motion History Histogram (HMHH) improves MHI representation [132]. However, these improved MHI-based representations at the low-level do not guarantee full description of action in the scene, therefore feature descriptors are used to characterize them in a representation ideally invariant to appearance and occlusions, background clutter, and even to shapes. As already mentioned by Ahad that no method, including

template-based method, can single-handedly address problems related to action recognition. This implies that a combination of relevant approaches, methods, techniques or representations are to be employed for better action recognition.

In a basic MHI-based approach [81], feature vectors representing Hu moments are calculated from both MHI and MEI images. Though these Hu moments can discriminate between shapes, they are difficult to reason about intuitively [81] and their computation is expensive depending of the template resolutions. Other approaches that take advantage of the textural riches of template such as MHI are found in the literature and are often referred to as textural feature-based approaches. These approaches make use of hand-crafted descriptors to examine each pixel and its neighborhood pixels, then extract relevant features through an encoding process. Necessary skills from feature engineering, which insure that resulting feature vectors from the process are discriminative and of low dimension, are required. Some of the commonly used feature descriptors for HAR are Histogram of Oriented Gradients (HOG) [16, 33, 55, 133–136] and Local Binary Patterns (LBP) [8, 68, 137–139] descriptors.

Histogram of oriented gradients [55] is initially used for object detection in computer vision and image processing. This descriptor counts the occurrences of gradient orientation in the localized ROI. Characteristics also known as HOG features that are derived from this approach correspond to the orientation histogram formed by considering both magnitude and direction computed at each pixel location using the gradient images. For a 2D grayscale image $I(x, y)$, gradient images G_x and G_y representing the horizontal and vertical gradient respectively are calculated by filtering the image $I(x, y)$ with the horizontal kernel $[-1, 0, 1]$ and the vertical kernel $[-1, 0, 1]^T$. From the computed gradient images, the magnitude is calculated as

$$m(x, y) = \sqrt{G_x^2(x, y) + G_y^2(x, y)}, \tag{2.14}$$

and the direction or orientation as

$$\theta(x, y) = tan^{-1}(\frac{G_x(x, y)}{G_y(x, y)}). \tag{2.15}$$

Equations 2.14 and 2.15 stand for the HOG feature extraction phase which also includes orientation binning, block description, and block normalization as processes yielding to the feature representation. The gradient image is then divided into cells and the orientation $\theta(x, y)$ obtained at each cell is quantized into orientation bins which are weighted by its respective magnitude $m(x, y)$ in order to form the corresponding histogram representing a block of a predefined number of cells. As many blocks of grouped cells are considered, they may be described by various geometries (e.g. rectangular R-HOG or

circular C-HOG) to overcome both illumination and contrast variations. The final HOG descriptor is a feature vector formed by concatenating the histograms of all the blocks after they have been normalized. These HOG features are usually fed to learning models such as KNN and SVM classifiers for recognition purpose. Many HAR applications have used HOG features to describe MHI templates. These HOG features were found robust as compared to the MHI descriptor when used uniquely [140]. And this because of the ability of HOG descriptor to capture local appearance and shape as a distribution of local intensity gradients or edge directions [16].

In [33], a multi-view HAR approach based on the description of MHI templates was proposed to overcome the high-dimensional representation due to multi-camera data at hand. Single MHI was considered at each viewed action video and later translated into HOG features, then classified using a simple nearest neighbor classifier. Good accuracies were also reported when MHI-HOG scheme was applied on well known benchmark Multicamera Human Action Video (MuHAVi) datasets [141]. The importance of using HOG description of MHI templates was also highlighted as the obtained results in terms of recognition rate outperformed the state-of-the-art approaches [140, 142]. However, these approaches can be computational expensive because they require setting up of multicamera during both training and testing phases as the person observed may be outside the range of camera. Therefore single-view approaches, though challenging, remain relevant in HAR applications and are presented in this thesis.

In [16], pedestrian activities are classified using motion patterns and HOG descriptors. The authors also highlight the impact of the HOG description in capturing both structure and shape within temporal templates such as MHI and MEI. Three schemes [16]; the MHI-HOG-SVM (MHS) model, MHI-MEI-SVM (MMHS) model and MHI-HOG-KNN (MHK) model, are compared. The MHS model with an accuracy of about 91% outperforms the MMHS and MHK models whose accuracies are 83% and 74% respectively. In addition, recognition rates obtained from different classifiers reveal HOG descriptors and SVM classifier as a better combination for action recognition. Similar conclusion is also reached in [143] as the authors compare performance of HOG-SVM, HOG-KNN, STIP-SVM, and STIP-KNN schemes through their validation on two popular datasets; KTH and Weizmann datasets. Besides the HOG-SVM scheme outperforming the HOG-KNN scheme, both schemes demonstrate that HOG features as a holistic feature representation is more discriminative than the STIP features, a local feature representation. Recognition accuracies provided by both HOG-KNN and HOG-SVM models are about twice those of STIP-KNN and STIP-SVM models, and also the time required to extract STIP features is about thrice that of extracting HOG features. However, this performance highlighted in [143] corresponds to features extracted from video frames and can be improved considerably if one uses a compact feature representation such as MHI

template per video. Ahad et al. [107] introduce the LBP operator as a texture operator, which encodes the direction of motion from non-monotonous areas of MHI templates, in combination with HOG descriptor to model human actions by describing human motions as texture patterns. This approach, that includes the SVM classifier, yields more accurate result with moderate computational expenses as LBP and HOG descriptors are computational simple feature extractors. Their scheme also symbolized by MHI-LBP-HOG-SVM (MLHS) is tested on the KTH action dataset and Pedestrian action dataset. Although LBP features have played a complementary role in the HOG-LBP feature representation, they have been proven to be more accurate, sparser, computational simpler than the gradient-based features such as HOG features [144, 145].

Originally introduced by Ojala et al. [41], the LBP operator is proven to be a powerful technique for texture description. Applications such as background modeling [146], face recognition [147], FER [148–150] and HAR [8, 68, 137–139] have gained from LBP and related variants such as CS-LBP, uniform LBP, LBP-TOP, VLBP, etc. for their tolerance in illumination changes, computational simplicity and speed. This LBP operator labels the pixels of a considered image by thresholding the local neighborhood around each pixel with the center pixel value resulting in a binary code. The formation of a basic LBP operator in the case of 3×3 neighborhoods is illustrated in Figure 2.15. The

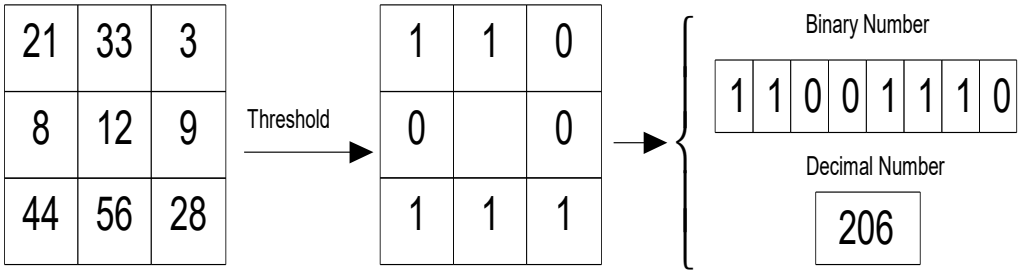


FIGURE 2.15: The basic LBP operator

resulting binary number is made of a total of eight bits yielding a total of 256 possible binary patterns. The basic of LBP operator is later extended to arbitrary circular local neighborhoods instead of a squared one as shown in Figure 2.15 to somehow achieve multi-scale analysis and rotation invariance [151]. The circular neighborhood is defined by a set of sampling points evenly spaced on a circle centered at the pixel to be labeled. The variables governing the circular local neighborhood structure are the number of sampling points and the radius of the circle, both denoted by P and R respectively. In addition, the sampling point not falling in the center of a pixel is chosen by means of bilinear interpolation. An example of the circular LBP operator is illustrated in Figure 2.16. Moreover, the theory of circular LBP operator is formally defined as follows

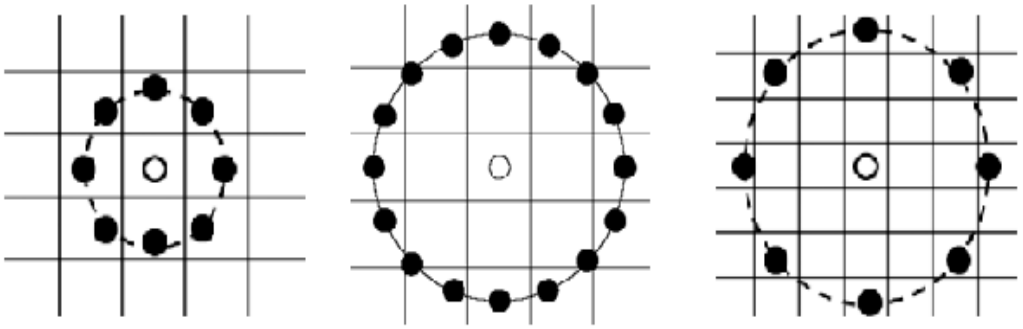


FIGURE 2.16: The circular (8,1), (16,2) and (8,3) neighborhoods respectively

[145, 151, 152]: Given a pixel at location (x_c, y_c) and its circular neighborhood (P, R), sampling points locations (x_k, y_k) are computed as

$$(x_k, y_k) = \left(x_c + Rcos\left(\frac{2\pi k}{P}\right), y_c - Rsin\left(\frac{2\pi k}{P}\right)\right), \tag{2.16}$$

and the intensity values at these sampling points denoted by $v_p = I(x_k, y_k)$ with $p = \{0, 1, 2, \ldots, P-1\}$. The basic LBP code is then expressed in the decimal format as:

$$LBP_{P,R}(x_k, y_k) = \sum_{p=0}^{P-1} S(v_p, v_c)\, 2^p, \tag{2.17}$$

where v_p and v_c are respectively intensity values of the center pixel and the p^{th} neighborhood pixels in the circular neighborhood (P, R), and the thresholding function $S(u)$ is defined as:

$$S(u) = \begin{cases} 1, & u \geq 0 \\ 0, & u < 0 \end{cases}. \tag{2.18}$$

For P spaced sampling points, the 2^P LBP codes ranging between 0 and 2^{P-1} are derived to form a LBP feature vector also identified as a histogram that is a collection of occurrences of different binary patterns representing different types of texture primitives such as edges, corners,flat areas, spots and lines ass shown in Figure 2.17. The above

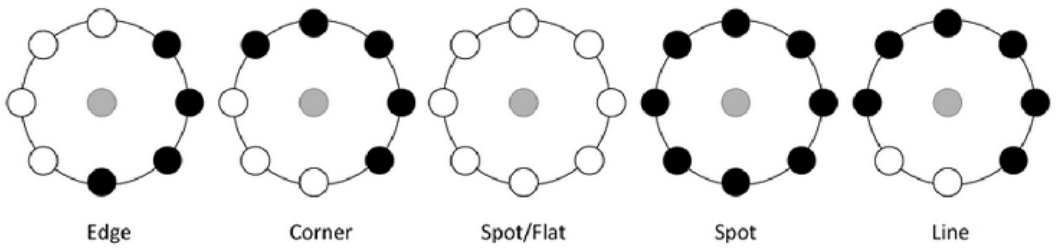


FIGURE 2.17: Examples of texture primitives detectable by LBP

definitions (Equations 2.16, 2.17 and 2.18) mark the feature extraction phase by expressing the properties of the LBP operator, which include its resistance to illumination variations and its simplicity in computation, making it attractive for FER [10] and HAR [8, 137–139, 153, 154] applications that share a common LBP feature-based methodology. In the literature, texture-based description of human action can be grouped in two categories:

- The first category is made of static texture-based method [8] that use temporal templates as still images produced at the preprocessing stage to characterize motion by extracting LBP histograms from both MEI and MHI templates. Based on the definition of each templates, it is observed that LBP patterns derived from the MHI template encode information about the direction of the motion whereas those derived from the MEI describe the overall pose and shape of motion. The global feature histogram representing an action is built by concatenating the MHI and MEI based LBP sub-region histograms. Figure 2.18 illustrates the formation of the LBP histograms used in sequential development of HMMs for recognition purpose.

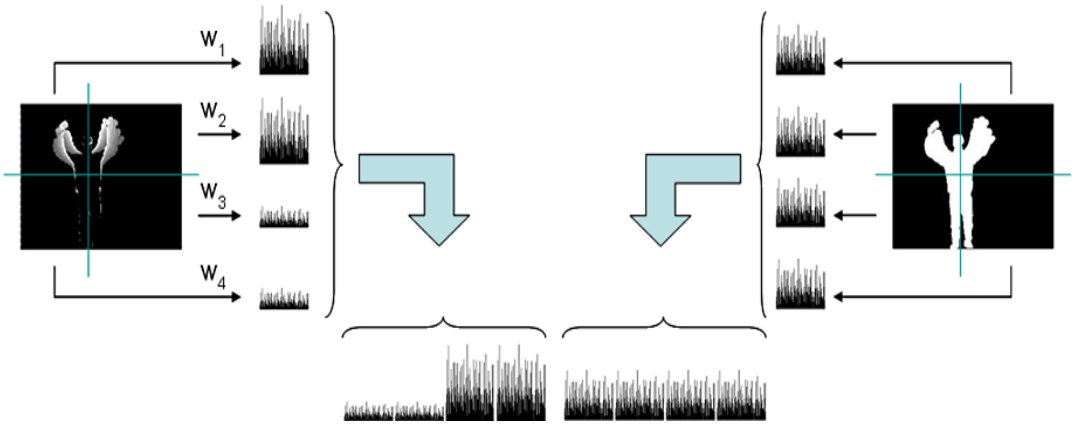


FIGURE 2.18: Illustration of the formation of a static based feature descriptor[8]

- The second category groups dynamic texture-based methods which are applied directly on video frames to extract dynamic texture features. In [9], the volume local binary pattern (VLBP) operator, an extension of the LBP operator, is used to model dynamic texture features as a texture descriptor combining appearance and motion. While operating in the spatio-temporal domain, this operator provides temporal features that are robust with respect to grayscale changes and geometric transformations such as translation and rotation. The VLBP operator, also called 3D-LBP, considers neighborhood frames to develop neighborhood sets in volume, as illustrated in Figure 2.19, to characterize the spatial structure of the local

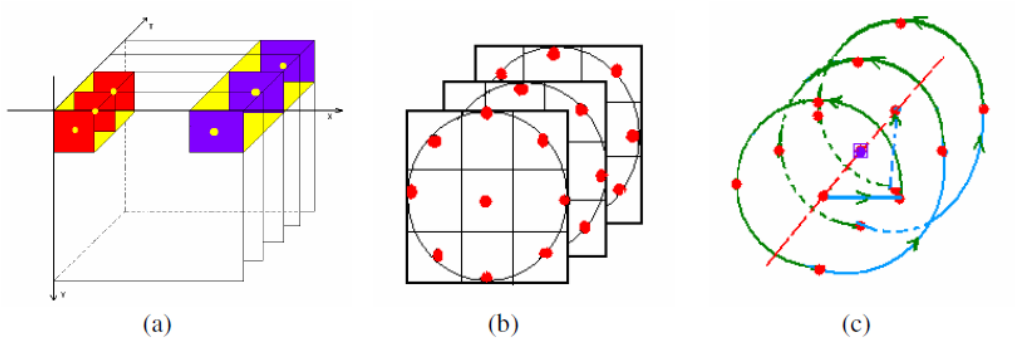


FIGURE 2.19: (a) Spatio-temporal volumes (Red formed volume for $L = 1$ and purple formed volume for $L = 2$). (b) Circular symmetric neighborhood sets in volume for $R = 1, P = 8$. (c) Neighborhood sampling point along the helix on the surface of cylinder for $P = 4$ [9]

volume dynamic texture as

$$VLBP_{L,P,R}(x_k, y_k, t_k) = \sum_{p=0}^{3P+1} v_p 2^p, \tag{2.19}$$

where (x_k, y_k, t_k) is the pixel location that locates the sampling points within the volume, L is the added dimension that represents the time interval and v_p is the joint distribution of grayscale levels that defines the dynamic texture in the local neighborhood. Although good accuracies with the VLBP dynamic texture features is reported for the MIT and DynText texture databases [9, 10], and also for Cohn-Kanade facial expression database [10], the number of patterns constituting the VLBP histogram is very large with a size of about 2^{3P+2} with respect to the increase of the P neighborhood sampling points within the volume. A simplified description of dynamic textures with the LBP-TOP operator is also proposed and compared to the description provided by the VLBP operator. The LBP-TOP operator describes dynamic texture in a spatio-temporal domain by concatenating LBP features extracted from three orthogonal planes as illustrated in Figure 2.20.

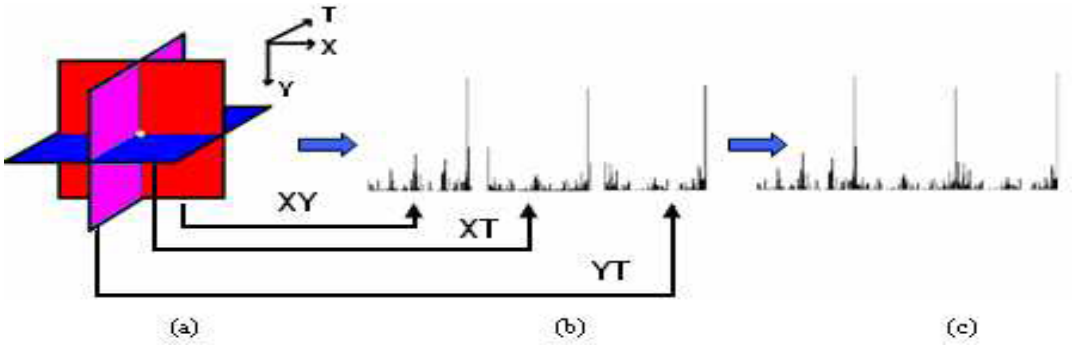


FIGURE 2.20: (a) Three orthogonal planes representing the dynamic texture. (b) Each plane LBP histogram. (c) LBP-TOP histogram as concatenated LBP histograms [10]

While reducing the computational complexity, this proposition lowers the length of feature vector to only 3×2^P as compared to 2^{3P+1} obtained with the VLBP operator. Besides LBP-TOP good performance on FER database, LBP-TOP has recently been applied in action recognition. In [138], the authors exploit the information about local or temporal locations by first dividing the bounding volume through its center into four sub-volumes. Thus the global feature histogram is built by concatenating single normalized sub-volume histograms as depicted in Figure 2.21. This histogram construction that we name here *histogramming* is a feature representation that improves the discriminative power of the texture features that plays a vital role in the performance during pattern recognition. Good recognition

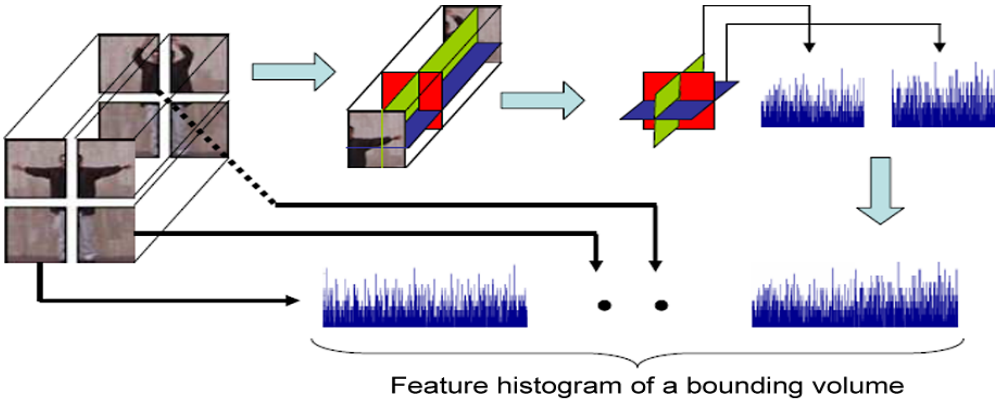


FIGURE 2.21: Global feature histogram formed from a bounding volume [10]

rates are reached with the Weizmann and KTH datasets despites a slight increase in length of the final histogram representing actions. However, histogramming as a technique for feature representation often yields to a phenomenon called the curse of dimensionality, especially when histograms computed from resulting blocks of a divided image or volume are concatenated for better recognition. This concatenation results in high dimensional feature vectors. For instance if one uses a circular $(8, 1)$ neighborhood to describe an image divided into 64 blocks, then the resulting $LBP_{8,1}$-feature vector is of high dimension equals to $16,384$. This phenomenon can also occur when considering a large amount of sampling points. So high dimensionality of feature descriptors implies high discriminative but low classification effectiveness in terms of speed both at the training and testing stages. Therefore, a high discriminative feature descriptor with low feature dimension is a requirement for HAR, especially in circumstances where timely responses are needed. For this reason, the computer vision community still seeks to re-engineer existing feature extraction techniques such as LBP descriptor or develop new ones for effective action recognition.

2.4.2 Deep learning feature-based methods

Appropriate and efficient representation of raw action videos principally influences the performance of HAR methods. Other methods, categorized as DL feature-based method and contrarily to handcrafted feature-based methods where action is represented by handcrafted feature descriptors, possess the capability to learn features directly or automatically from raw data (images or videos). The concept of end-to-end learning is then introduced implying transformation from the pixel level to the action recognition. Resulting trainable feature extractors and computational models of multiple processing layers are used for action representation and recognition. This type of method, due to the existence of advanced technology nowadays, is also referred to as modern approach and now leads the state-of-the-art researches in the field of computer vision. These methods are also known as deep neural networks and rely on the adaptation of multilayered neural deep architectures to process real-world data [14]. So depending on how architectures are built and what applications, e.g. generation or recognition, it is intended for, Li Deng and Dong Yu [155] broadly categorize most of the DL models into three major classes: (1) Supervised or Discriminative models such as Deep Neural Networks (DNNs), Recurrent Neural Networks (RNNs), and Convolutional Neural Networks (CNNs); (2) unsupervised or generative learning models such as Deep Boltzmann Machines (DBMs) and Deep Belief Networks (DBNs); (3) Hybrid models, which are models that exploit the characteristics of models in category (1) and category (2) mentioned above.

2.4.2.1 Discriminative models

The most used discriminative models found in the literature for HAR under the supervised class are the Convolutional Neural Network (CNN) models. Because of their ability to extract hierarchical features which are built by executing convolutional operations alternating with subsampling operations [11, 156], deep learning models have shown superior performance at various computer vision tasks, such as object detection [157], facial expression recognition [158], human pose estimation [159], and human action recognition [13, 14, 160–163]. CNN models, while processing real-world data [14], rely on the adaptation of multilayered neural deep architectures in the form of multiple hidden layers to transform bulky raw inputs into class outputs. Their architectures are generally composed of three principal types of layers; convolution layer, pooling layer also known as subsampling layer, and full connected layer. An example of a simple CNN model used to recognize characters is depicted in Figure 2.22. In the convolutional layer, various kernels are utilized to convolve the entire considered input data as well as the intermediate feature maps to generate various new feature maps. Doing so, the convolution operation presents few benefits which are; the weight sharing mechanism in

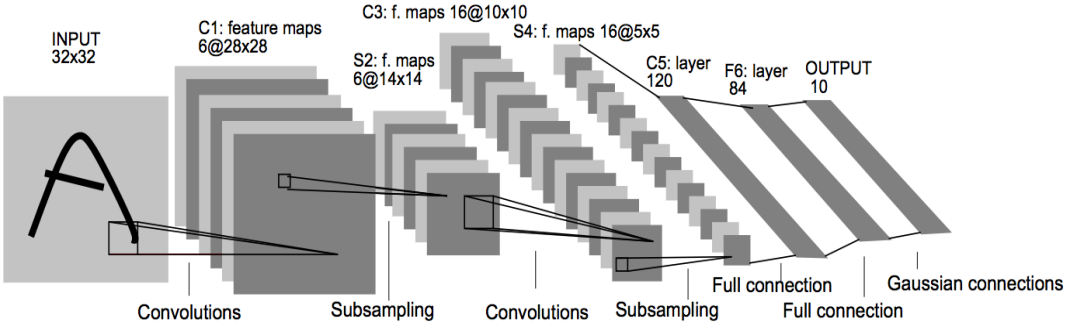


FIGURE 2.22: Basic architecture of CNN model dubbed LeNet-5 by LeCun et al[11]

the same feature maps reducing the number of parameters, the local connectivity learning correlations among neighboring pixels as it is done partly when computing LBPs, and finally the invariance to object location within the convolving input data. These benefits also justify the choice of CNNs as a substitute for fully connected layers in the traditional neural networks, which speed up the training process. Following the convolutional layer, the pooling layer aims at reducing the size of the previous feature maps for the next convolutional layer. The pooling operation, also referred to as subsampling or downsampling, usually does not affect the depth of the information although some loss of information is experienced in the process. This loss is advantageous since the reduction of size yields less computational overhead for the next layer within the network. Most CNN architecture are wrapped up by a couple of fully connected layers which are added after several combined convolution and pooling layers. FC layers are where the high reasoning levels are performed. Here feature maps are converted into a 1D feature vector which is later fed into a classifier; e.g softmax [164].

Moreover, the architecture of CNNs embodies three major design characteristics which include the local receptive field, tied or shared weights, and subsampling [165]. These characteristics, while describing the operation of CNN models, contribute to strong robustness to both scale and translation variations of local features, and also to the reduction of trainable parameters. Based on local receptive field, a set of neighboring units belonging to the previous layer constitutes inputs to each unit in the next convolution layer. By convolving these inputs with convolution filter or kernel as illustrated for a 2D input image in Figure 2.23, neurons are then given the capability to extract elementary visual features such as oriented edges, corners, endpoints, etc., forming a feature map. The feature map, also referred to output feature map, is constructed from input feature map through a convolution operation, followed by an additive bias and an activation

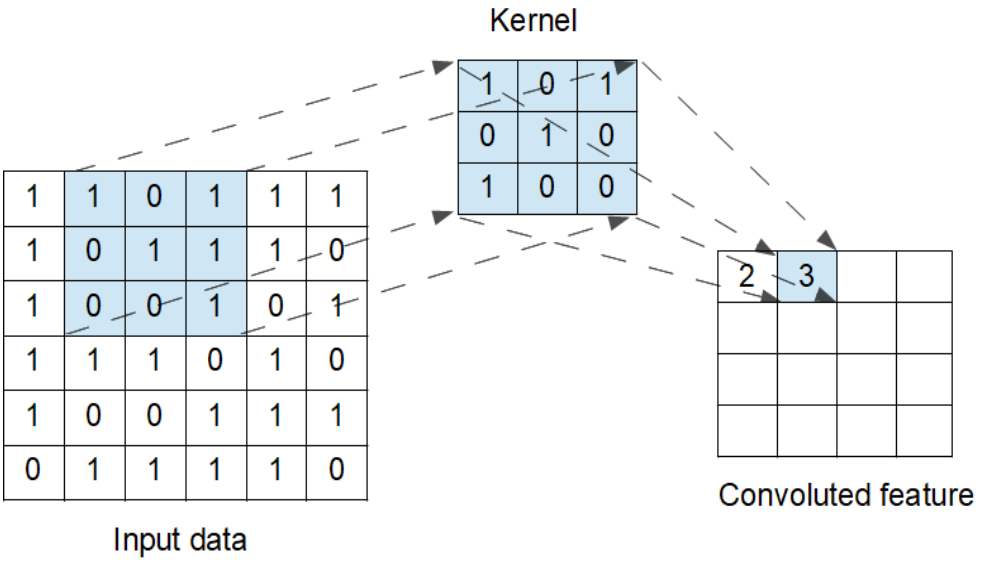


FIGURE 2.23: A typical convolution operation

function, defined mathematically by

$$y_j^l = f\left(\sum_{i \in M_j} y_i^{l-1} * k_{ij}^l + b_{ij}^l\right), \tag{2.20}$$

where y_j^l is the output of the current layer, y_j^{l-1} is the previous layer output, k_{ij}^l is the kernel for the current layer, b_j^l are the biases for the current layer, and M_j is a selection of input feature maps. The function f can be either linear or nonlinear activation function such as sigmoid, hyperbolic tangent, rectified linear, or identity functions. Among the nonlinear activation functions as shown in Figure 2.24, the rectified linear function also known as Rectified Linear Unit (ReLU) is found better function over the sigmoid and hyperbolic tangent functions mostly used in traditional neural networks because ReLU presents benefits such as alleviation of gradient problems, sparse activation, unilateral inhibition, and wide excitation boundary. Without the introduction of activation

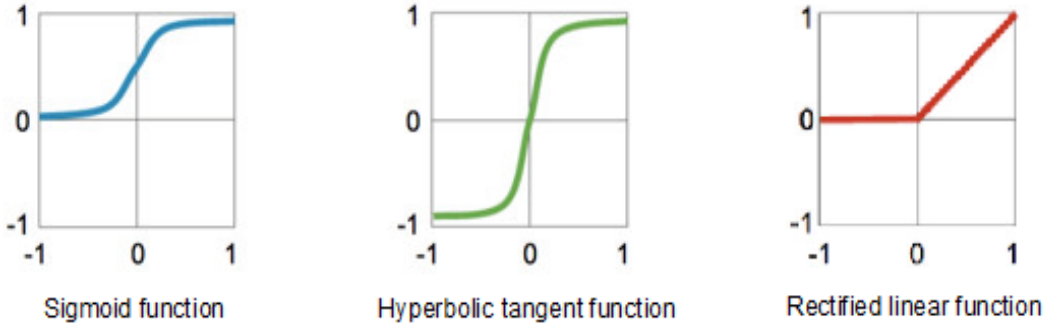


FIGURE 2.24: Activation functions: Sigmoid, Hyperbolic tangent, and Rectified linear function

function, ReLU for instant, in-between or after sets of convolution and pooling layers, a CNN model cannot be deepened indefinitely, which prevents the aggregation of lower order features. The combination of these features through subsequent convolution layers produces higher order features which are more discriminative. The idea of using tied weights, also referred to parameter sharing, reduces significantly the number of parameters to be learned and ensures hard-coded invariance to translation or position into the architecture. Such organization of weights also reduces the storage requirements of the CNN model and does not affect the runtime of the forward propagation [166]. The third characteristic referring to the subsampling transforms the convolution output at a certain location into a summary statistic of the nearby outputs [166]. This transformation can be defined as

$$y_i^l = sub\left(y_i^{l-1}\right), \tag{2.21}$$

where $sub(.)$ represents a subsampling function. The max pooling and the average pooling are the two most used subsampling techniques and are illustrated in Figure. 2.25. The ability of the max pooling to select superior invariant features, improve generaliza-

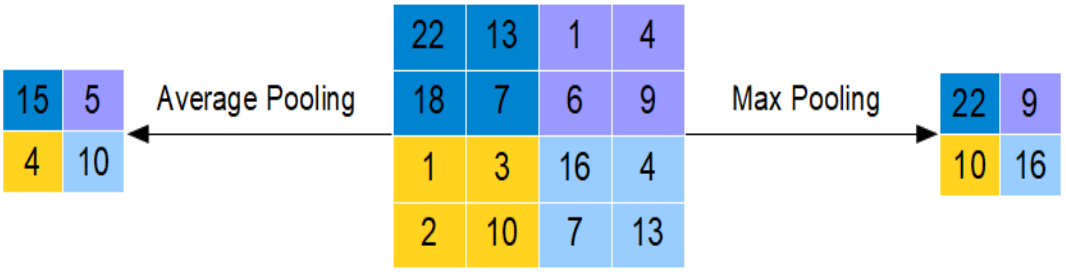


FIGURE 2.25: Average versus Max Pooling

tion, and lead to fast convergence is demonstrated in [167]. This feature of CNN model, while playing an important role in the CNN performance, assists as well in controlling the total parameter count . However, the efficiency of a CNN model is enhanced through the optimization of its output feature map which is highly dependent on hyperparameters such as kernel size, depth, stride, and zero-padding. As these hyperparameters ditacte the spatial arrangement, they are also used to calculate the size of the output of a convolutional layer as

$$O = \frac{V - F + 2P}{S} + 1, \tag{2.22}$$

where V is the input volume size (height×width×depth), F is the receptive field size, S is the stride, and P corresponds to the amount of zero-padding used on border and set to be

$$P = \frac{F - S}{2}. \tag{2.23}$$

In addition, each hyperparameter has a key role in the performance of a CNN model;

the kernel size decides on the number of learnable weights which if lower yields computational efficiency, the output depth controls neurons count which if reduced significantly minimizes the total number of neurons within the network, the stride configures how far the filter or kernel slides or spans within the given input volume during convolution operations, and lastly the application of zero-padding forces the size for the output of the convolutional layer to be idem to the input one. Besides the above-mentioned hyperparameters, excluding the zero-padding which is not contributing to the reduction of the dimension of the feature vectors when the CNN model is used as feature extractor. Dropout [168, 169] as a regularization technique is often used to alleviate the problem of overfitting during training process while giving a more efficient representation of the model. However, the tuning of hyperparameters and the use of Dropout through optimization algorithms such as gradient descent (GD) [170], root mean square propagation (RMSProp) [171], adaptive moment estimation (Adam) [172], adaptive gradient (AdaGrad) [173] or stochastic gradient descent (SGD) [174] remains essential for any action recognition applications to cope with basic requirements such as speed, accuracy, and consistency.

Deep CNN models were initially used for representing and recognizing objects for still images [175, 176]. An illustration of a 2D-CNN model also called LeNet-5 [11] developed by LeCun in the early age of CNN is depicted in Figure 2.22. These 2D-CNN models generally handle 2D raw input data which, in a HAR application, are referred to as video frames. This type of model can only learn spatial information and ignore the motion information encoded in multiple adjacent video frames. This ignorance of the motion information was overcome by the use of stacked video frames as input to the deep network though obtained results underperformed when compared with the handcrafted shallow representations [177]. Later, through an investigation done by Simonyan and Zisserman to address the issue, a two-stream (spatial and temporal) convolutional neural network is proposed for action recognition [12]. Figure 2.26 shows the proposed two-stream CNN architecture where both the spatial stream and the temporal stream recognize the action from still video frames and from the motion in the form of dense optical flow respectively. Subsequently, features resulting from these streams are combined through fusion for action recognition purpose. Although this approach demonstrates superior results as compared to those obtained from the shallow handcrafted-based representation methods [178], it is still not suitable for real-time applications because of its computational complexity. Hence a need for a simple CNN model as a single stream, which could handle video sequence as a unique 3D space-time volume, was imposed. Ji et al. answered to this need by introducing a 3D CNN model as an extension of a 2D convolution to the time space while extracting both spatial and temporal (spatial-temporal) features simultaneously for action recognition. Figure 2.27 illustrates how multiple features are

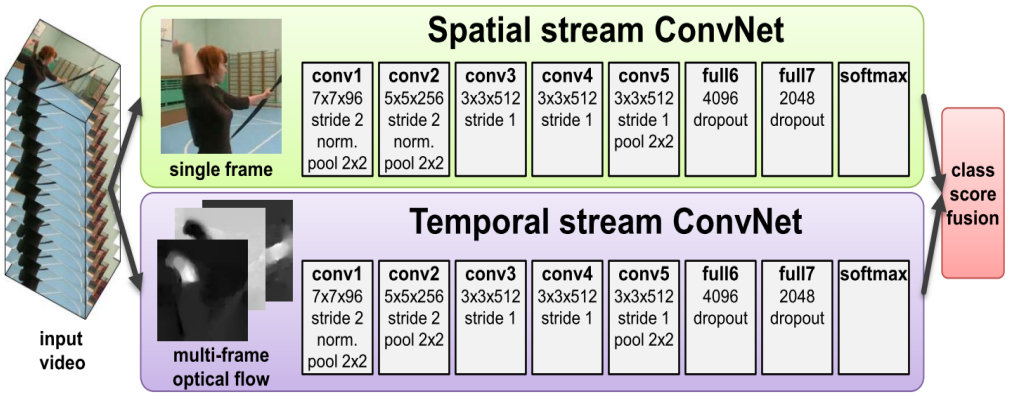


FIGURE 2.26: A two-stream CNN architecture for video classification[12]

extracted from adjacent frames. Kernels of 3D sizes were used for 3D convolutions in

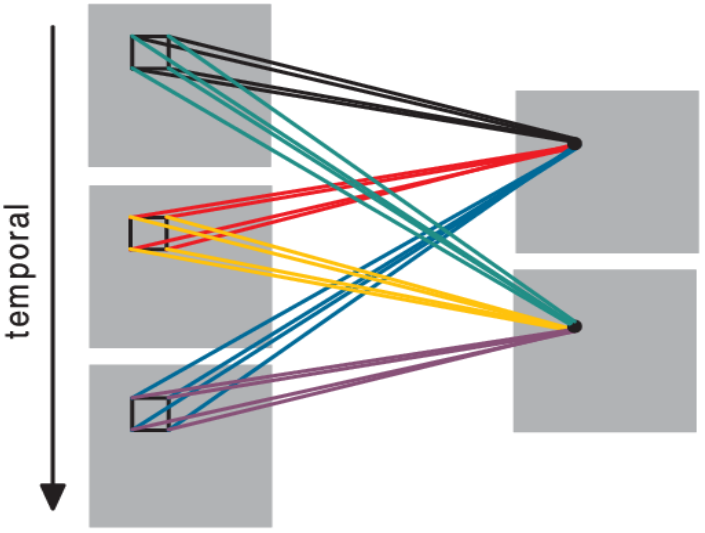


FIGURE 2.27: Illustration of the extraction of multiple feature from multiple 3D convolutions applied to adjacent frames [13]

convolution layers to abstract spatial-temporal information in a natural way at multiple semantic level from videos. The proposed 3D CNN architecture consisted of a hard-wired layer that encodes prior information in the form of grayscale, gradient, dense optical flow frames by dividing the raw input into sets of 3D CNN layers and spatial pooling layers, which were then recombined in a succeeding FC layer feeding a softmax output layer with six separate outputs matching each output to an action class. Motion features were learned by the 3D CNN layers which are part of the proposed 3D CNN model depicted in Figure 2.28. A competitive average error rate of 9.8% across a 5-folds of the data was achieved when applied to KTH action dataset. Although a remarkable success of CNN models is highlighted in the literature, their application requires huge amount of labeled data and suffers of high computational complexity during the training of convolutional kernels. Some works have been proposed to overcome these problems. In [179], Sun

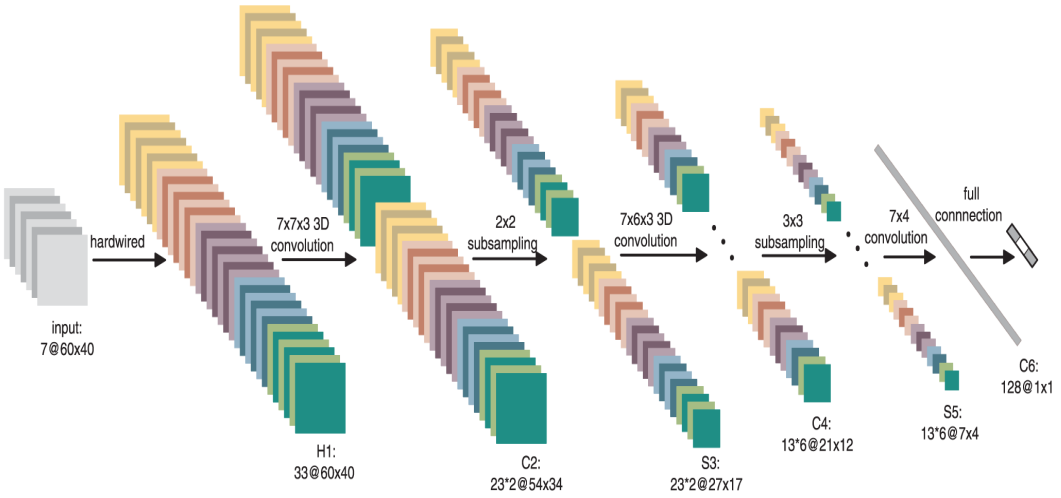
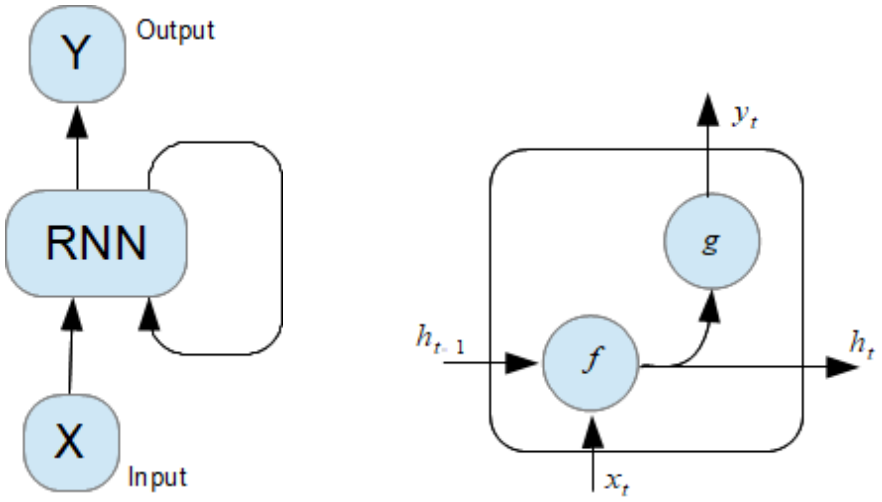


FIGURE 2.28: Illustration of a deep 3CNN model developed by Ji et al [13] for HAR. The color-coded sets of connections illustrate shared weights in similar color.

et al. proposed factorized spatio-temporal convolutional networks (FSTCN) for HAR which addressed the computational complexity by simplifying the standard 3D CNN model into a 2D spatial kernel a lower layers guided by the sequential process and 1D temporal kernels in the upper layers. This approach reduced the number of parameters which evenly cut down the computational complexity of training process of CNN kernels. Among the 3D CNN models [14, 180–182] used for HAR, the 3D CNN model by Tran et al. [183], which also exploited spatio-temporal features, was considered as the deeper network and validated on four public datasets confirming the suitability of 3D CNN models over the 2D CNN, the size $3 \times 3 \times 3$ as the best kernel size for spatio-temporal features, and the better performance of 3D CNN with linear classifier over the state-of-the-art approaches. However, these 3D CNN models are usually learned within a short snippet of the video and fail to model actions over their full temporal extent [182, 184]. The size of convolutional kernels and the number of considered video frames as input to the CNN model contribute to this failure as they restrict the captured range of dependencies between video frames. Consequently, typical 3D CNN models, though providing competitive accuracies, are not easily adaptable to long range dependencies. Varol et al. [182] propose long-term temporal convolutions (LTC) for action recognition that improves the recognition accuracies proportionally with respect to the increase of sampled video frames to the expenses of costly runtimes.

Furthermore, recurrent neural network (RNN) is another deep learning approach specialized for sequential learning modeling. This approach is also a supervised method that models complex dynamics of various actions in videos by means of its architecture which by its cyclic connections as shown in Figure 2.29(a) allows to both store and access the long range contextual information of temporal sequences. Figure 2.29(b) illustrates the basic function of a RNN node where h_t and x_t are respectively the hidden state and



(a) RNN Block diagram (b) RNN node structure

FIGURE 2.29: Basic RNN architectures: (a) Block diagram of a RNN model consisting of a cyclical connection of RNN nodes handling temporal sequences; (b) Single RNN node structure describing its internal operation.

the output at current time t and computed based on the previous hidden state h_{t-1} and the input x_t at current time by the following equations:

$$h_t = f\left(W_h h_{t-1} + U_h x_t + b_h\right) \tag{2.24}$$

$$y_t = g\left(W_y h_t + b_y\right), \tag{2.25}$$

where U_h, W_h, and W_y are the weights for input-to-hidden recurrent link, hidden-to-hidden link, and hidden-to-output link, respectively. The bias terms b_h and b_y correspond to the hidden and output states, respectively. Similarly to CNN models, f and g are the activation functions associated with each RNN node and can be any of the functions illustrated in Figure 2.24. For the reason that RNN models are effective in the capture of temporal information as current prediction is function of both current observation and memorized information in hidden states, they have to a great degree been applied in action recognition to model human motion in videos and produced improved performance over CNN models. Recurrent neural networks as initially introduced by Hochreiter and Schmidhuber [185] know the vanishing gradient problem which makes them very difficult to train [186]. To address this problem, a variant of RNN called long short-term memory (LSTM) method, which allows for constant error signal propagation through time, was proposed in [185]. The LSTM architecture, equipped with an internal cell state and three recurrent gates; input gate, forget gate, and output gate, allows its memory cell to store and access temporal information over an extended period. Contrarily to typical RNN model, LSTM model with its capability to model temporal sequences and their wide-range dependencies more accurately [187] were proven suitable

for action recognition. In [14], the authors tackle the question of recognizing human actions by combining two networks; CNN and LSTM, forming a CNN-LSTM model (see Figure 2.30). Thus LSTM network is fed with output from a 3D-CNN network in order

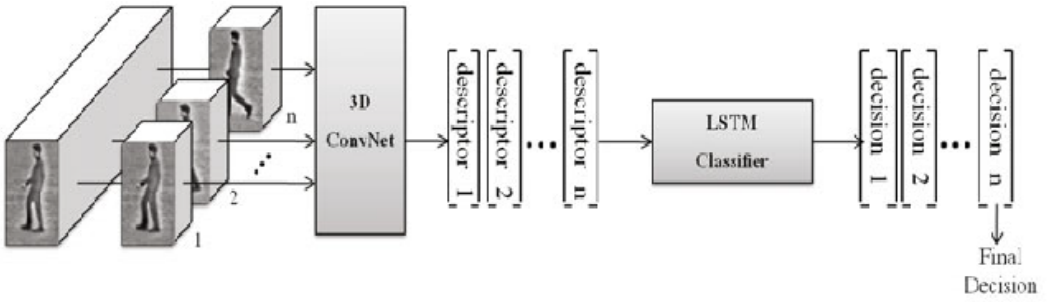


FIGURE 2.30: Illustration of a deep CNN-LSTM model developed by Baccouche et al [14] for HAR.

to classify the entire sequence using learned features. This combination as applied to KTH dataset by Baccouche et al. [14] yielded a reduction of the average error rate to 5.61% outperforming the error rate obtained when only a 3D-CNN model is considered. This scheme was recently adopted in the works of Danahue et al. [188], Ng et al. [189], Li et al [190], and Chen et al [191]. Besides using LSTM as a classifier, it has also shown its effectiveness as an end-to-end framework [192, 193] for action recognition purpose. However, LSTM based methods are made out of huge amount of parameters per unit which unveils their computational complexity.

2.4.2.2 Deep generative and hybrid models

Applying backpropagation algorithms to deep networks with many layers forms part of problems addressed by deep learning models, especially when they should not only learn the nonlinear mapping between input and output data but also the underlying structure of the input data [194] . Then unsupervised pretraining is required to recognize unlabeled action datasets. This approach, also known as unsupervised or generative deep learning model, uses RBMs or auto-encoders in each hidden layer forming the model. Two examples of unsupervised learning model for action recognition; DBN and DBM models, were noticed in the literature. These deep generative models belongs in the Boltzmann family since they utilized the RBM as learning module, which is a generative stochastic neural network allowing for more efficient training algorithms such as the gradient-based contrastive divergence algorithm [195]. DBN models, being graphical models and learning to extract deep hierarchical representations of training data, were used to learned invariant spatio-temporal features from video [196–199]. As for DBM models, their architecture is similar to that of the DBN models at the exception that all connections at the layers are undirected [200, 201]. Despite its architecture that makes

them a complete hierarchical generalization of RBM with the ability to capture many layers of complex representations, these models (DBM) have not yet been applied to visual based action recognition. This may be unfortunately due to learning procedure that remains too expensive for large-scale computer vision problems. However, it is worth mentioning that generative models such DBN and DBM models are predominantly used for wearable sensor-based applications, and have been overshadowed by supervised learning models due to the practicability of CNN in deep learning.

Furthermore, hybrid model, also referred to as a combination of generative and discriminative models or different models within the same category, have been implemented in action recognition by many researchers. For instance where generative and discriminative models are combined, Chen [202] proposed a spatio-temporal DBN architecture stacking spatio-temporal convolutional RBM layers for the extraction of ST features. Although this architecture owns the attribute of generative model by nature, it nevertheless possesses convolution elements making it be referred to as a hybrid model. In addition, this model did not achieve state-of-the-art results on KTH action dataset when compared to handcrafted feature based methods, but shown superior performance when compared to its counterparts, convolutional deep belief networks (CDBNs) [202]. Additional information on combinations of CNNs with DBNs forming CDBNs can be found in [203–205]. As for the combination of models within the same category, hybrid models built from combining discriminative models are very much visible in the literature. Thus the use of CNNs with LSTMs has shown some state-of-the-art improvement in many popular datasets [14, 188–191]. Here the CNN part is used to learn both effective and robust features directly from raw video data, and the LSTM part is used to learn statistical dependencies over contiguous actions in order to recognize action sequences. A softmax classifier often forms the final layer for action classification purpose. Such CNN-LSTM models have improved the state-of-the-art results in many benchmark datasets. However, more models imply an increase in the number of parameter leading to an increase of the computational complexity.

2.5 Classifiers for human action recognition

HAR algorithms are applied on visual data e.g. image or video, to extract discriminative and salient features Various classifiers are used to complete the recognition process for the purpose of recognizing action classes. This recognition is done through action representations which can be either a handcrafted action representation or a deep learning action recognition. Depending on the type of action representation, an appropriate classifier is used. The state-of-the-art classifiers are KNN and SVM classifier for handcrafted

feature based approaches, and Softmax for deep learning approaches. In this section we discuss these classifiers by emphasizing on their particularities with respect the fields of machine learning and deep learning.

2.5.1 K-nearest neighbor (KNN) classifier

The KNN classifier, as one of the simplest machine learning algorithms [206], classifies objects based on most similar training samples in the feature space. It requires no training model to be built before the recognition stage. Here the algorithm locates the k nearest training samples to a given query test sample and determines its class label by selecting the single most frequent class label of the nearest training samples. The k nearest neighbors are selected accordingly to the k smallest distances which are calculated based on some distance metric such as Euclidean distance or Hamming distance. The class label with the majority voting on the k nearest neighbors is then assigned to the class label of the training sample. In addition to its simplicity and ease of implementation, the KNN algorithm has produced state-of-the-art results in HAR applications [207, 208]. However, this classifier presents some limitation which are; large memory requirement for a large training set, sensitivity to the choice of similarity function, sensitivity to irrelevant features, and degradation of prediction accuracy with growth of attributes. Consequently, a good performance of this classifier is then expected when low dimensional feature descriptors representing objects or actions are used in recognition stage.

2.5.2 Support vector machines (SVM) classifier

A SVM classifier is a supervised learning technique which, based on a designed set of hyperplanes in a multi-dimensional space that separates data points from different classes with a large margin [209], solves the classification problems. Data points closer to the separating hyperplane define the support vectors. Usually this learning model distinguishes between two classes, which is not the case for action recognition problem where more that two actions has to be recognized. Figure 2.31 illustrates how a distinction between two classes is made by a SVM. So a multi-class SVM adopting a one-against-all approach is required. Each SVM aims then to discriminate between an action and all the other remaining actions. Another particularity to SVM is that it deals effectively with high dimensional feature descriptors that are generally extracted from handcrafted feature based techniques. This classifier has gained popularity among visual-based recognition [17] because of good results obtained when learning from handcrafted features.

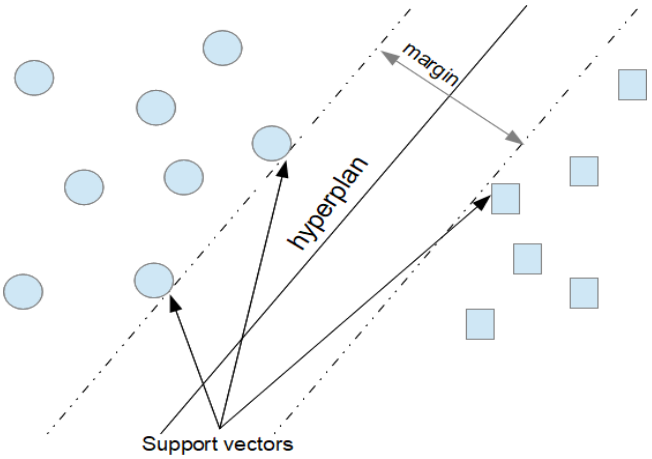


FIGURE 2.31: SVM classification for two classes

2.5.3 Softmax classifier

Softmax classifier is quite popular in the context of deep learning and convolutional neural networks. This classifier constituting the last layer of DL models, is used to compute the probabilities for each class label. The class label with the highest probability is then considered as predicted class label for the query input data. Mathematically, the softmax function also referred to as an activation function $f : \mathbb{R}^J \to \mathbb{R}^J$ is defined by

$$f(\mathbf{x})_i = \frac{e^{x_i}}{\sum_{j=1}^{J} e^{x_j}}, \tag{2.26}$$

for $i = 1, \ldots, J$ and $\mathbf{x} = \{x_1, \ldots, x_J\} \in \mathbb{R}^J$. In classification tasks, J is the number of classes corresponding also to the number of neurons of the output layer. The output of the Softmax also describes a probability distribution since it does not only maps the output to the range $[0, 1]$, but maps each output in such a way that the sum of all the outputs equals to 1. Besides its probabilistic properties, this function can be used for multi-classification contrarily to other activation functions which are more indicated for binary classification. For instance, in HAR classification recognition problem, one can interpret $f(\mathbf{x})_i$ as the estimated probability of the network that the correct action classification is i defined as

$$i = \arg \max_{i=1,\ldots,J} f(\mathbf{x})_i \,. \tag{2.27}$$

Moreover, this classifier is a very simple model, an effective predictor, and very fast to train. It has been the most favorable used classifier in discriminative [176, 179–182, 188–193], generative [194, 195], and hybrid [203, 204] deep learning models.

2.6 Datasets for action recognition

Datasets play a important role on the approval of the effectiveness of both old or newly developed HAR approaches. These datasets should present all possible challenges whose few are highlighted in Figure 1.1 and discussed in Section 1.1. This section briefly describes three publicly available benchmark dataset that have paid remarkable contribution in HAR research. These selected datasets are all visual based data and are used to validate various contributions made in this thesis.

2.6.1 JAFFE Facial Expression Dataset

The first dataset is the JAFFE dataset [15] and has a total of 213 still images of seven facial expressions whose six basic facial expressions; sad, happy, disgust, fear, surprise and angry, and one neutral. This database was planned and assembled by Lyons et al. [15] and the facial images of Japanese female models taken in the frontal head pose was rated by 60 Japanese subjects. The original size of each image is 256×256 pixels. A sample set of images from JAFFE database is depicted in Figure 2.32.

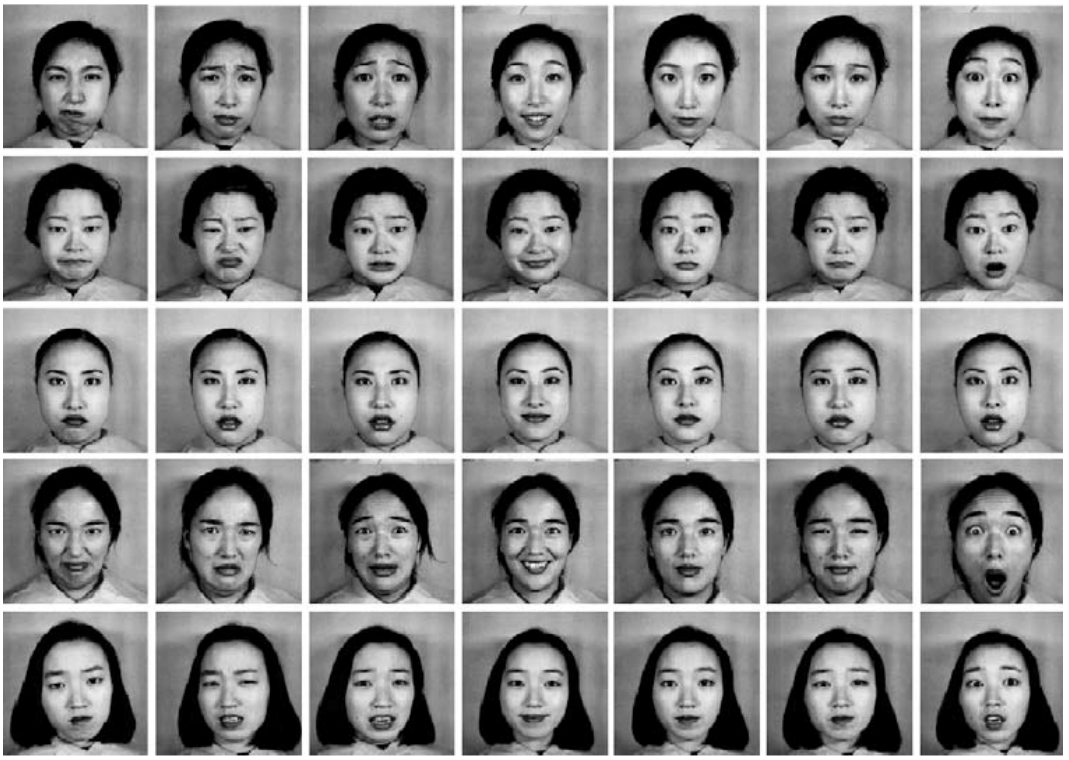


FIGURE 2.32: Sample images from JAFFE dataset [15]

2.6.2 Pedestrian Action Dataset

The Pedestrian action dataset consists of a wide range of possible pedestrian actions taking place while crossing a road. This database, which initially used in [16], was also the object of the evaluation of many other methods [107, 210, 211] proposed for action recognition. Pedestrian actions such as walking or running straight from one side to the other side of the road, turning towards or opposite to the camera, turning and returning to where pedestrian started from, cross walking, falling down accidentally and cannot rise up, falling down and rising up immediately, were recorded in an indoor environment with a Nikon D3200 camera from a stationary position. Figure 2.33 shows some example of Pedestrian action dataset. Moreover, the dataset presented a total of 160 video clips

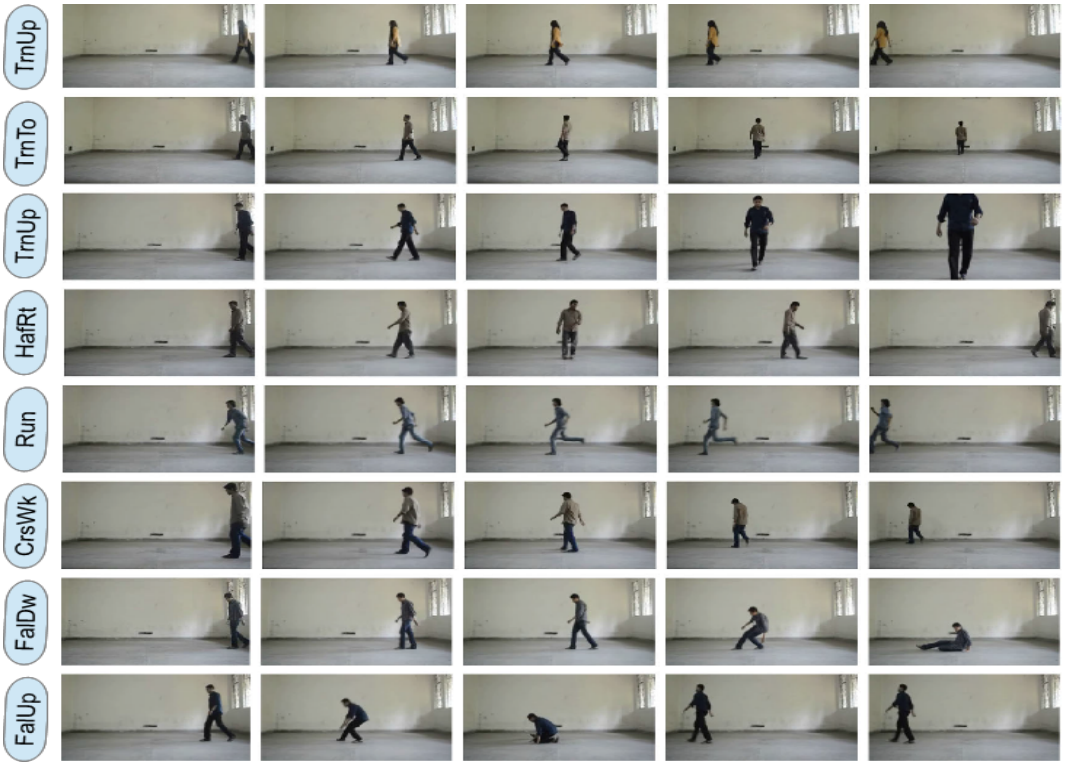


FIGURE 2.33: Sample frames from Pedestrian action dataset [16]

scaled at a resolution of 170 × 320 of twenty persons performing each eight different actions. Table 2.1 provides a detailed description of the Pedestrian action dataset.

2.6.3 KTH Human Action Dataset

The KTH action dataset is one of the mostly used benchmark dataset for HAR that was introduced by Schuldt et al. [17]. This dataset consists of 25 subjects performing six different human actions: Boxing, Handwaving, Handclapping, Jogging, Running, and

TABLE 2.1: Detailed description of Pedestrian Action Dataset [16]

Action label	Description
Walk	walk straight from one side of the road to the other side of the road with the camera on side view of the pedestrian
Run	run straight from one side of the road to the other side of the road with the camera on side view of the pedestrian
TrnTo	turn towards the camera in which case the camera gets at first the side view and then the frontal view of the pedestrian
TrnOp	turn opposite to the camera in which case the camera gets at first the side view and then the back view of the pedestrian
HafRt	turn and return to where the pedestrian started from
CrsWk	cross walk
FalDw	fall down accidentally which may be serious to the instant that the pedestrian does not rise up
FalUp	fall down accidentally which may not be serious to the instant that the pedestrian rise up

Walking. The video sequences were recorded in four different scenarios: outdoors, outdoors with scale variations, outdoors with different clothes, and indoors. The database contains a total of 2391 sequences which were taken over homogeneous backgrounds with static camera with a frame rate of 25 fps, downsampled to a spatial resolution of 160 x 120 pixels with an average length of four seconds. Sample frames of the KTH dataset are shown in Figure 2.34.

FIGURE 2.34: Sample frames from KTH action dataset [17]

2.7 Summary

This chapter presented a comprehensive literature review of previous methods along with theoretical tools used by researchers to tackle the problem of recognizing human actions over the past twenty years. These methods, which could be either local or holistic, were split according to how they performed action representation between handcrafted feature based approaches and deep learning feature based approaches. As both methods were informed by the type of visual data at hand, spatio-temporal features related to action recognition were then discussed in Section 2.2. It was therefore found that visual data, either in the form of image or video, were not beneficial for HAR in their raw state, which gave room for new, compact and discriminative spatio-temporal features. In Section 2.3, the local action representation was briefly discussed. This representation was produced from local methods which encoded video sequences as an accumulation of local spatio-temporal features referred as local descriptors. Although these local methods would prove useful in the literature, their collection of only independent patches through the detection of STIPs and the description of these local patches around the interest points contributed to their lack of interpretability and also to expensive computational costs.

The state-of-the-art of the holistic action representation, which could be based on either handcrafted features or deep learning features, because of its simplicity and good interpretability of actions was also discussed in Section 2.4. Handcrafted methods were found to be either statistical or structural feature based. The statistical based holistic approaches employed analyses such as PCA, LDA, FLDA, and RDA acting as dimensionality reduction techniques to recognize human actions. The PCA was initially used but due to its lack of separability inability the LDA was proposed. This LDA, also known as a generalization of FLDA, could recognize actions but suffered of both SSS problem and the null space which could actually possess discriminative properties during feature representation. These problems were later addressed by the RDA approach which affected the parametric eigenspectrum modeling and sub-space decomposition with the aim of minimizing the classification error. Many RDA variants which improved the recognition performance were proposed despite the high computational complexity caused by the search of best-fit parameters. Consequently, a need to explore structural based holistic approaches was then imposed.

Further in Section 2.4.1, precisely in Section 2.4.1.2 structural based holistic approaches with respect to silhouette, motion and spatio-temporal templates were discussed in details. These approaches were defined as techniques that represented human actions by their global appearance and motion while preserving their structural information made of both spatial and temporal structure of action taking place in a video sequence. Silhouette-based approaches as a first attempt to structural based holistic approaches,

characterized actions in video by capturing shape features from human image or silhouette (Figure 2.6) from background extraction. In addition to their subjectivity to noise due to an fallible background extraction, they could not capture motion information, a key role player in the description of videos. Thus motion-based approaches captured the motion feature between successive video frames and its success in extracting good descriptors depended on reliable optical flow estimation algorithms which were found to be computational expensive. A little further in Section 2.4.1.2, another structural based holistic approaches which considered both the spatial and motion information and known as spatial-temporal based approaches were then discussed. These approaches originated from accumulated human silhouettes resulting in compact forms called spatio-temporal motion templates or simply temporal templates which at their turn provided global features as global action representation. These templates included MEI, MHI, MHV (an extension of MHI in 3D), and related variants. Among these templates, more attention was paid to MHI templates whose employability with feature descriptors such as HOG and LBP shown good results in applications such as action recognition. These descriptors as feature extraction techniques were used to characterize actions in a representation ideally invariant to appearance, occlusions, background clutter, and shapes; and the length of resulting feature vectors had a great impact on the computational burden of classifiers such as KNN and SVM. However handcrafted features with high discrimination power and very low dimensionality for action representation will ever be a requirement for high performance in terms of accuracy, speed and consistency.

Contrarily to handcrafted feature based methods, DL feature based methods that could learn features directly or automatically from raw data were discussed in Section 2.4.2. These methods were divided into three groups of models: discriminative, generative and hybrid models. These models represented deep neural networks which relied on the adaptation of multilayered neural deep architectures to process data. CNN models were found as the most used discriminative models in the literature for HAR applications followed by LSTM models which served for sequential learning modeling. The combination of CNN and LSTM models helped to deal with both spatial and temporal information within a video and yielded a better recognition rate. Other architectures comprising either CNN, LSTM, or both were also presented in this chapter. Further deep generative and hybrid models were briefly discussed. The former known as an unsupervised models used RBMs in each hidden layers forming their architectures. Two models; DBN and DBM models, were identified. Besides DBN ability to learn invariant spatio-temporal features from videos and DBM having not been applied to visual based action recognition, both models have been predominantly used in wearable sensor-based applications. The latter which referred to a combination of both generative and discriminative models were implemented in action recognition but did not achieve state-of-the-art results.

However the problem of computational complexity remained a challenge since good recognition results would come with a huge number of parameters from either deeper or combined DL models.

Action representation in the form of feature vectors derived from either handcrafted or deep learning feature-based methods were fed to classifiers to recognize the action label from a given query video. KNN, SVM and Softmax classifiers were then identified as the most used in holistic HAR systems and were discussed in Sections 2.4.1.2 and 2.5. KNN and SVM classifiers were found suitable for Handcrafted approaches whereas the Softmax classifier for DL approaches. The outperformance of SVM over KNN classifier was noticed in comparisons made among different schemes applied to KTH, Weizmann, and Pedestrian action datasets.

Lastly, in Section 2.6 well-known public datasets such as JAFFE, KTH and Pedestrian action datasets were presented for evaluation purpose. Motivated by the fact that most of HAR algorithms applied to videos are usually an extension of algorithms applied to 2D still images and that FER is associated to HAR application, the JAFFE dataset was elaborated in a proposed feature descriptor in Chapter 3.

In closing, as reported earlier in this chapter, holistic methods for action representation for HAR are practically divided into handcrafted and deep learning feature based approaches. The handcrafted feature based methods also known as traditional methods rely on descriptor algorithms with the ability to produce suitable feature vectors from reduced database for good performance, whereas the deep learning ones known as modern methods rely on their deep architectures trained with sufficiently large database on powerful hardware to generate robust feature representations for good performance. Both methods are limited by their high computational complexity which remains a great concern for real-time applications. To address this concern, a new feature descriptor as a variant of the very popular LBP descriptor because of its low computational complexity and two other existing global feature descriptors are proposed in a handcrafted fashion for the recognition of human actions in Chapter 3. In Chapter 4, the simplest form of the 2D-CNN model preceded by a preprocessing stage is proposed to lower the computational complexity.

Chapter 3

Contributions to handcrafted feature based approach to human action recognition

"Thinking is easy, acting is difficult, but bringing thoughts into action is the hardest thing in the world."

Johann Wolfgang Goethe

3.1 Introduction

After reviewing the literature that forms the base of handcrafted feature based approaches in Chapter 2, we realize that the issue of computational complexity in HAR can be addressed by the use of simple descriptors producing handcrafted features with very low dimension. In this chapter, we thus propose the use of existing global feature descriptors in a framework manner and a new feature descriptor for action representation. These descriptors extract handcrafted features from MHI template which is widely used for human action recognition (HAR) due to its simple representation of the motion information to recognize human actions. However, the interpretation of human activities in video sequences remains a challenging task and their recognition is not yet mastered in the field of computer vision. Ahad [106] states that action recognition and behavior analysis are still in their infancy even though they have been investigated for few decades. Consequently there is still room for performance improvement since the ability to recognize actions relies more on their suitable feature descriptors. Few handcrafted feature based approaches are presented in this chapter.

Firstly, a method that includes a holistic feature extraction technique not yet employed in HAR applications, named slope pattern spectra (SPS) is proposed in Section 3.2. Increasing slope pattern spectra are extracted from motion history images and fed into a K-Nearest Neighbor (KNN) classifier in order to recognize various human actions. The proposed framework is later tested on the KTH dataset, a mostly used benchmark dataset for HAR. The experimental results demonstrate that SPS are suitable feature descriptor for HAR via MHI.

Secondly, Section 3.3 presents a novel feature extraction technique called circular derivative local binary pattern (CD-LBP). Motivated by uniform local binary patterns (uLBPs) which exhibits high discriminative potential at a reduced data dimension of the original LBP feature vector, CD-LBP feature descriptors are created as a result of binary derivatives of the circular binary patterns formed by LBPs. The description through CD-LBP can provide lower dimension and high discriminative feature vectors as handcrafted features yielding to improved recognition rate at a minimum running time. Higher order CD-LBP descriptors are the lower dimensional feature descriptors which give flexibility in deciding on recognition system performance, especially if a real-time implementation is considered. Two experiments are used to validate the proposed CD-LBP algorithm. JAFFE dataset and KNN classifier are used in the first experiment and the Pedestrian action dataset and SVM classifier used in the second experiment to validate the proposed CD-LBP algorithm. The HOG descriptor is used in combination with CD-LBP descriptor to represent pedestrian actions. However, results in both experiments demonstrate the relevance of the proposed CD-LBP feature description especially when performance metrics such as recognition accuracy and running time value are considered.

3.2 Slope pattern spectra for human action recognition

Motion history image (MHI) method, also referred to as a smart action representation approach by Ahad [106], has gained popularity due to its simple motion representation of a video sequence into a single grayscale image. In this MHI image, dominant motion patterns are preserved. Despite the fact that MHI method records the history of temporal changes at each pixel location, it still suffers from the problem of self-occlusion for certain actions. Most HAR-related application has used MHI in combination with other techniques to extract more motion information for better results. One common limitation when employing only MHI method or its variants is the size of feature vectors which is still large and uncomfortable for action classification. Based on the idea that if a proper feature extraction technique is applied to MHI images, MHI is sufficient to describe the human action in video sequence. Thus the proposed method includes the

extraction of slope pattern spectra (SPS) [43] from MHI images. The use of SPS is motivated by the fact that a MHI image while describing motion occurence, a spread of temporal changes at each pixel location develops a slope progression which can be either in an increasing or decreasing order. The increasing slope pattern spectra as handcrafted features are then extracted from MHI images derived from the KTH Action dataset and classified using a KNN classifier to demonstrate the importance of SPS applied to MHI for HAR. The proposed framework which includes MHI generation, SPS extraction, KNN classification is discussed in Section 3.2.1.

3.2.1 Proposed method

The proposed method aims at improving the recognition performance of MHI-based human action recognition. The distribution of slope patterns developed in the MHI templates are extracted using the SPS algorithm. To the best of our knowledge, the application of slope pattern spectra (SPS) to HAR has not been proposed before. We are motivated by the fact that this approach is a holistic feature extraction technique, and is used to extract global motion information from MHI. The proposed method comprises two main parts: feature extraction standing for action representation and classification. Our proposed methodology is illustrated in Figure 4.1.

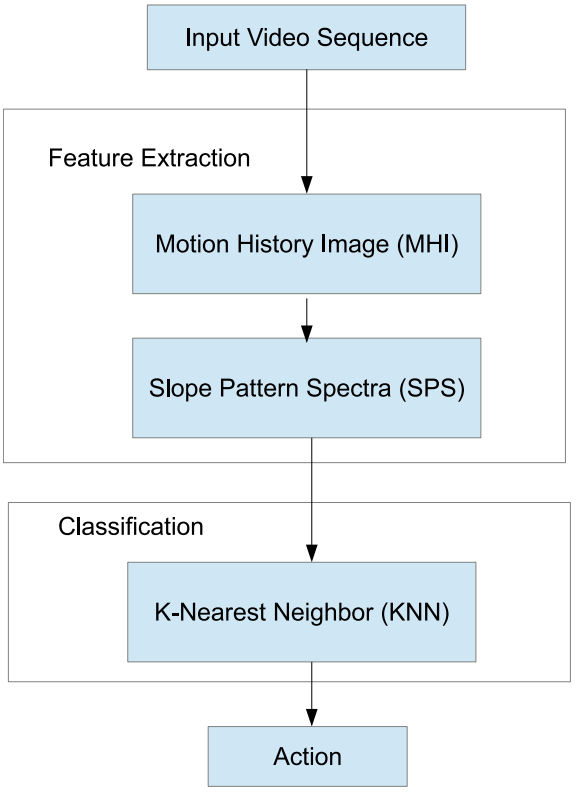


FIGURE 3.1: Proposed methodology.

Motion History Image– MHI templates are generated by keeping history of temporal changes at each pixel location, which then decays over time [212]. The resulting MHI is an image in which grayscale value indicates the most recent motion of the pixel in a set of video sequences. The brighter intensity values of the MHI correspond to more recent motion [213]. The concept of MHI has been overstretched in [107] and defined by Equations 2.11, 2.12 and 2.13 recalled here as follows:

$$H_\tau(x,y,t) = \begin{cases} \tau & \text{if } \Psi(x,y,t) = 1 \\ max(0, H_\tau(x,y,t) - \delta) & otherwise \end{cases} \tag{3.1}$$

where $\Psi(x,y,t)$, the update function is defined by

$$\Psi(x,y,t) = \begin{cases} 1 & \text{if } D(x,y,t) \leq \epsilon \\ 0 & otherwise \end{cases} \tag{3.2}$$

and $D(x,y,t)$, the difference of frames is given by

$$D(x,y,t) = |I(x,y,t) - I(x,y,t \pm \lambda))| \tag{3.3}$$

and $I(x,y,t)$ is the intensity value at the pixel location (x,y) of the t^{th}-frame in the video sequence. The parameters τ, δ, λ, and ϵ whose effects on generated MHI templates are well described in Section 2.4.1.2 must be adjusted carefully for good motion representation. Although they have an inexpensive computation, MHI templates still experience a problem of noisy background and occlusion for some human actions where motion information overwrite. Some examples of MHI image are shown in Figure 3.2.

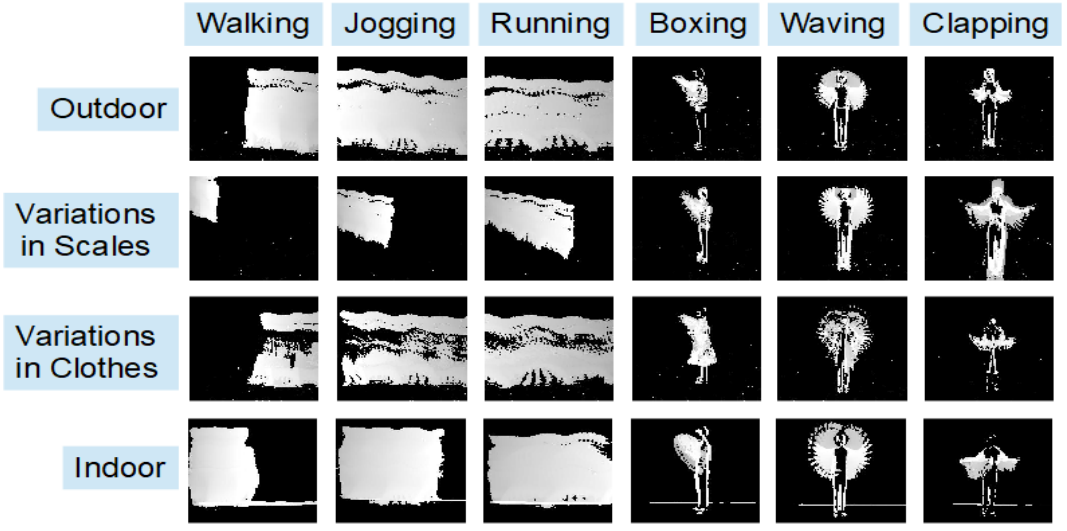


FIGURE 3.2: Examples of MHI template corresponding to each scenario.

Slope Pattern Spectra– To address the occlusion problem, SPS algorithm for global feature extraction is used. This algorithm is classified among global feature extraction techniques for texture analysis. The SPS algorithm was developed in [43] and applied to seed mixture, High Steel Low Alloy (HSLA) steel, and satellite images to extract their corresponding global image signature. This signature, as pattern spectrum or feature descriptor represents the distribution of formed increasing slope segment (ISS) in an image. The algorithm is also computationally inexpensive and described as follows [43]: Consider a grayscale image I where horizontal lines of I (rows of I) are taken one after the other, and scanned from left to right. In each horizontal line, possible slope segments (SS) are determined. If the slope segment is an increasing slope segment (ISS), the measure of the increasing slope segment divided by its length, $\frac{m(ISS)}{n}$, is calculated, and the n^{th} bin of the pattern spectra incremented with this value. If the lengths of the increasing slope segments determined at different horizontal lines are equal, the same pattern spectrum bins are respectively incremented by $\frac{m(ISS)}{n}$. These operations are repeated until the last horizontal line of I is processed. See Algorithm 1.

Algorithm 1: Slope Pattern Spectra algorithm [43]

Initialization:
Pattern spectrum: for all $n>0$, $PS[n]\leftarrow 0$
Calculation:
for each horizontal line (row) of the grayscale image f
Initial integral pixel value: $F(0)\leftarrow 0$
Length of ISS : $l(ISS)\leftarrow 0$
$\Delta\tilde{F}\leftarrow 0$
for each pixel value p_i in the horizontal line with $i\geq 1$ do:
$F(i)\leftarrow F(i-1)+p_i$ i.e. the integral image;
$\Delta F(i)\leftarrow F(i)-F(i-1)$ i.e. the integral segment derivative;
if $\Delta F(i)\geq\Delta\tilde{F}$
$l(ISS)\leftarrow l(ISS)+1$ i.e increase length of the ISS ;
$\Delta\tilde{F}\leftarrow\Delta F(i)$;
else
$n\leftarrow l(ISS)$ i.e. determine the length of the ISS ;
$m(ISS)=F(i-1)-F(i-n-1)$ i.e. determine the measure of ISS ;
$PS[n]\leftarrow PS[n]+\frac{m(ISS)}{n}$ i.e. add the contribution of ISS to n^{th} bin of the SPS ;
$\Delta\tilde{F}\leftarrow 0$;
$l(ISS)\leftarrow 0$;
end if
end for
end for.

As the SPS descriptor capitalizes on the distribution of motion patterns in the MHI template to extract suitable features for HAR, its derived feature size is minimum which therefore reduces the computation load on the classifier. Examples of increasing slope pattern spectrum are shown in Figure 3.3.

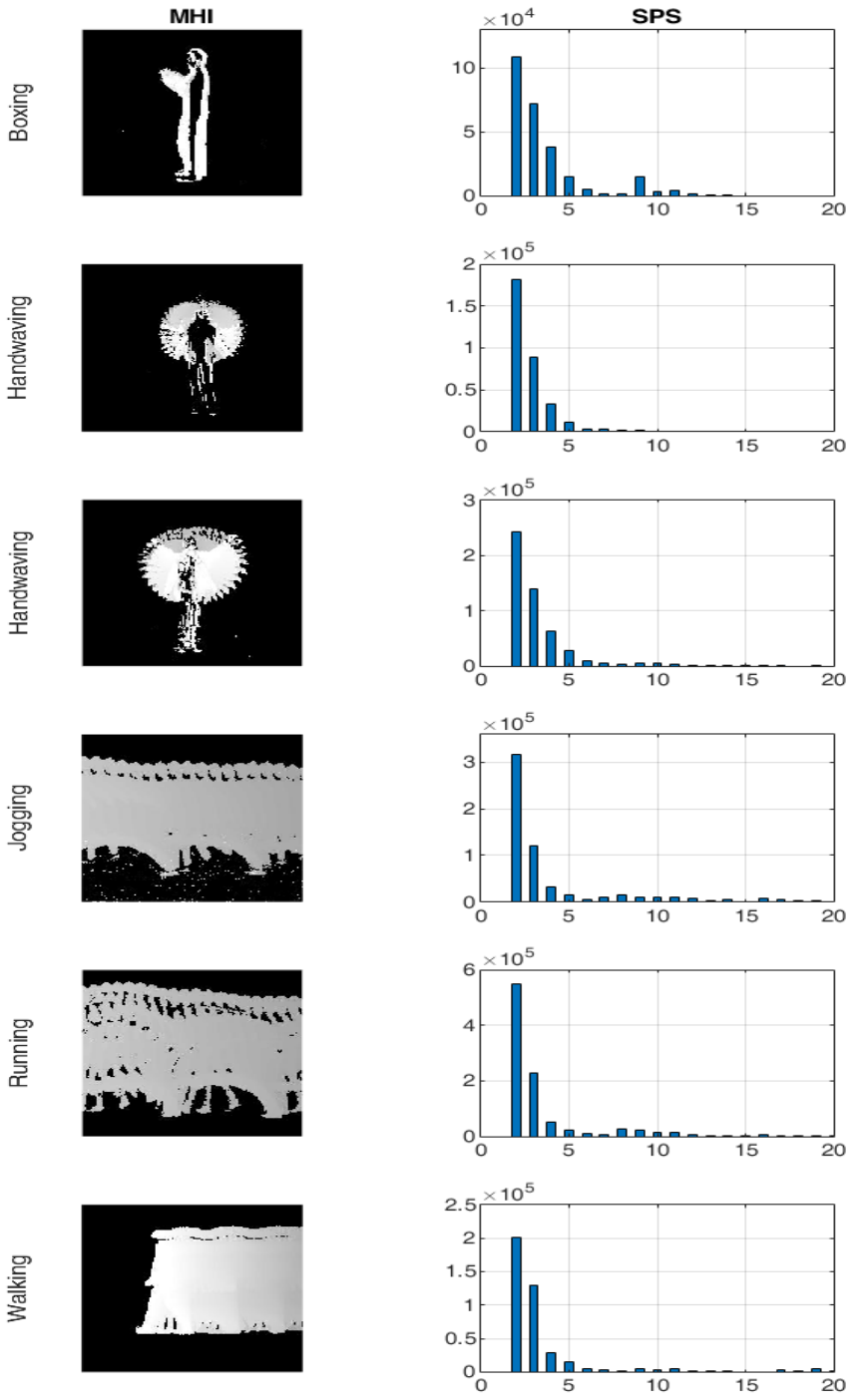


FIGURE 3.3: Examples of MHI and their SPS respectively.

K-Nearest Neighbor Classifier– The second and important part of our proposed method is the classification part which plays a role in deciding the predicted action based on its corresponding action representation. The KNN classifier is chosen because of the simplicity of its implementation compared to other learning machine algorithms [207, 208]. Moreover, the handcrafted features derived from the SPS algorithm are of low dimension and best fit the KNN classifier since no dimension reduction technique is required to avoid the classifier to suffer the curse of dimensionality. Other attributes that make the KNN classifier suitable for the proposed framework are discussed in Section 2.5.1.

3.2.2 Experimental results and discussions

An experiment is conducted on KTH action dataset [17] using the Matlab software to validate the proposed method. First, MHI images are generated from all video sequences without doing any partition in terms of training or testing sets. Second, each MHI are translated into SPS features along with respective action label. Last, the derived SPS features are partitioned and fed into the classifier. The results are then presented with respect to feature extraction and classification.

3.2.2.1 Feature Extraction result

The parameters for the computation of MHI templates are critical to capture relevant motion information. As suggested by Ahad [106], the following MHI parameters $(\tau, \epsilon, \delta) = (255, 40, 2)$ are considered. Fifty frames in total per video constitutes each video sequence which is then processed to generate the MHI image. Few MHI images are not too convincing as visually one can spot some kind of occlusion or noise, specially the MHI images obtained from the video subset taken outdoors with scale variation. Example of MHI with self-occlusion and noise are shown in Figure 3.4. Such MHI images have a costly effect on SPS feature extraction not leading to high accurate classification.

3.2.2.2 Classification results

The K-nearest neighbor classifier is used. The extrated SPS feature vectors are divided into two sets, consisting of 80% and 20% of 600 SPS feature vectors for the training and testing set respectively. Two classifications are conducted on the KTH Dataset due to the complexity presented by this dataset. In the first classification, action videos are grouped with respect to their conditions (outdoor (S1), outdoor with scale variation

(a) Running (b) Walking

FIGURE 3.4: Examples of MHI for outdoor video sequence.

(S2), outdoor with clothing variation (S3) and indoor(S4)). The second classification is for the overall KTH Dataset.

Per Scenario Classification results– Tables 3.1-3.4 use confusion matrix summarize classification results for each scenario.

TABLE 3.1: S1 confusion matrix: mean performance of 56.67%

Actions	Box	Clap	Wave	Jog	Run	Walk
Boxing (Box)	40.0	20.0	20.0	0.00	0.0	20.0
HandClapping (Clap)	16.7	66.7	0.0	0.00	0.0	16.7
HandWaving (Wave)	0.0	0.0	75.0	12.5	0.0	12.5
Jogging (Jog)	20.0	0.0	0.0	20.0	20.0	40.0
Running (Run)	0.0	0.0	0.0	0.0	100	0.0
Walking (Walk)	0.0	25.0	25.0	0.0	0.0	50.0

TABLE 3.2: S2 confusion matrix: mean performance of 50.00%

Actions	Box	Clap	Wave	Jog	Run	Walk
Boxing (Box)	55.6	11.1	22.2	0.0	0.0	11.1
HandClapping (Clap)	33.3	66.7	0.0	0.0	0.0	0.0
HandWaving (Wave)	0.0	0.0	33.3	33.3	0.0	33.3
Jogging (Jog)	25.0	50.0	0.0	0.0	0.0	25.0
Running (Run)	0.0	0.0	20.0	0.0	80.0	0.0
Walking (W)	0.0	16.7	16.7	16.7	0.0	50.0

TABLE 3.3: S3 confusion: matrix mean performance of 63.33%

Actions	Box	Clap	Wave	Jog	Run	Walk
Boxing (Box)	57.1	0.0	28.6	0.0	0.0	14.3
HandClapping (Clap)	25.0	25.0	25.0	0.0	0.0	25.0
HandWaving (Wave)	0.0	0.0	75.0	0.0	0.0	25.0
Jogging (Jog)	25.0	0.0	0.0	75.0	0.0	0.0
Running (Run)	0.0	0.0	0.0	28.6	71.4	0.0
Walking (Walk)	0.0	0.0	0.0	25.0	0.0	75.0

TABLE 3.4: S4 confusion:matrix mean performance of 60.00%

Actions	Box	Clap	Wave	Jog	Run	Walk
Boxing (Box)	50.0	16.7	16.7	0.0	0.0	16.7
HandClapping (Clap)	16.7	66.7	0.0	0.0	0.0	16.7
HandWaving (Wave)	20.0	0.0	80.0	0.0	0.0	0.0
Jogging (Jog)	0.0	0.0	0.0	100	0.0	0.0
Running (Run)	0.0	0.0	0.0	25.0	75.0	0.0
Walking (Walk)	20.0	0.0	0.0	60.0	0.0	20.0

Out of 150 SPS feature vectors, 120 and 30 are used for training and testing purpose respectively. A 100% classification rate per scenario is obtained for validation. The average performance for each scenario classification with respect to testing is computed as the trace of each matrix over the total number of classified elements (e.g. From Table 3.1, the mean performance is 50%). Other testing average performances are indicated on respective tables. These results are quite encouraging since they demonstrate that the SPS feature vectors can be discriminative and used for recognition purpose. The S2 classification performance is the lowest and can be justified by the scale variations in video sequences.

Overall Classification result– The overall classification result, which average is 50.83%, is illustrated in Table 3.5. This average overall performance is lower than the 63.50%

TABLE 3.5: S4 confusion matrix:mean performance of 50.83%

Actions	Box	Clap	Wave	Jog	Run	Walk
Boxing (Box)	31.6	42.1	26.3	0.0	0.0	0.0
HandClapping (Clap)	16.7	50.0	27.8	0.0	0.0	5.6
HandWaving (Wave)	14.3	19.0	66.7	0.0	0.0	0.0
Jogging (Jog)	0.0	0.0	0.0	35.0	30.0	35.0
Running (Run)	0.0	5.3	5.3	15.8	64.2	10.5
Walking (Walk)	4.5	13.6	9.1	13.6	0.0	59.1

found by Meng et al. [132] when using only the MHI as feature descriptors. Though the discrepancy of rates is of 13.50%, the length of the SPS feature vector is far less than the one of MHI, making classification computational less expensive.

In summary, the proposed method that includes the SPS algorithm as a feature extraction technique is applied for the first time in the field of HAR. The SPS handcrafted features are simply computed from the MHI images to form a low dimensional feature space that leads to a robust classification. The per scenario classification results indicate the complexity of the KTH Action dataset. Hence the obtained results are quite encouraging and demonstrate the relevance of slope pattern spectra as holistic action representation.

3.3 Circular derivative local binary pattern for human action recognition

Human action recognition remains a challenging task in the field of human behavior understanding, more especially when human actions must be recognized from static images. Although a static image contains less information when compared to a video, it is shown that the information in a single image is enough to achieve acceptable recognition rates [148, 214, 215]. The use of still images therefore requires well suited feature descriptors in order to extract relevant handcrafted features. Many texture descriptors have recently attracted great attention for the recognition of human actions, especially the LBP which has become very popular for both its computational simplicity and speed [147]. However, the LBP descriptor despite its high discriminative power slows down the running time of classifiers because of its feature size leading to lengthy feature histogram. The long feature set is due to the fact that LBP is derived from the adoption of only the first-order gradient information between the center pixel and its neighborhood [216], and also the concatenation of LBPs extracted from local regions of an image. Consequently, the development of a feature extraction technique that, regardless of variations in human subjects and capturing conditions, provides lower dimension and high discriminative feature vectors yielding to improved recognition rate at a minimum running time, is still an open question. By considering the recognition rate and running time as performance metrics, a feature extraction technique that does not compromise much on either of the performance metrics is required. In other word, a novel descriptor that can achieve a good trade-off between recognition accuracy and computational efficiency is essential for real-time applications. So here we propose a novel feature extraction technique called circular derivative local binary pattern (CD–LBP). This proposed descriptor, inspired by the uniform LBPs which are patterns with at most two circular 0-1 and 1-0 transitions [217], takes advantage of the circular binary structure of the LBP to evaluate the relation between consecutive binary digits. This relationship, also referred to as a circular derivative, allows the proposed CD–LBP descriptor to provide higher-order CD–LBP features as lower dimensional handcrafted features affording a flexibility in deciding on the performance of the recognition system, especially in a real-time implementation.

Furthermore, two applications of the CD–LBP algorithm as experiments are also presented here for validation purpose. JAFFE dataset and KNN classifier forms parts of the first experiment. For the second experiment,the HOG descriptor used in combination with CD-LBP descriptor to extract handcrafted features to represent pedestrian actions.

3.3.1 Circular derivative local binary pattern descriptor

Feature extraction is a very important phase of any pattern recognition system. The most important characteristics of a feature extraction technique is its ability to be highly discriminative and the low dimensionality of the feature vector produced. Such characteristics have high potential to yield good classification accuracies. The proposed descriptor builds on the basic concept of the LBP operator discussed in Chapter 2. It is also inspired by the uniform LBP which is an useful extension of the original LBP operator used to reduce the length of the feature vector.

Based on Equation 2.17 and looking into the generated $LBP_{P,R}$-patterns, two types of patterns are observed; the non-uniform patterns and the uniform patterns. The uniform patterns are found to carry more information than the non-uniform ones and are described as binary pattern with at most two bitwise transitions from 0 to 1 or vice versa. A total of $P(P-1)+2$ uniform patterns can be extracted using a circular (P,R) neighborhood. This feature extraction technique as an extension of the basic LBP descriptor is denoted by $LBP^{u2}_{P,R}$ and defined as

$$LBP^{u2}_{P,R} = \begin{cases} LBP_{P,R} & \text{, if} \quad U(LBP_{P,R}) \leq 2 \\ \\ P(P-1)+2 & \text{, otherwise} \end{cases} \tag{3.4}$$

where $U(LBP_{P,R})$, the uniformity measure is defined as

$$U(LBP_{P,R}) = |S(v_{P-1} - v_c) - S(v_0 - v_c)| + \sum_{k=1}^{P-1} |S(v_k - v_c) - S(v_{k-1} - v_c)| \tag{3.5}$$

with S being the thresholding function defined in Equation 2.18.

Considering a circular $(8,1)$ neighborhood, the uniform patterns account for nearly 90% of all the LBP patterns. Examples of uniform pattern are 11111111 with no transitions and 00111110 with 2 transitions. The pattern 110000101 is not an uniform pattern since it has 4 transitions. The $LBP_{8,1}$ has 256 possible patterns whereas the $LBP^{u2}_{8,1}$ only 59 possible patterns which is in fact 58 patterns added with one single pattern for all non-uniform patterns. The uLBP descriptor produces a low dimension feature set therefore will definitely yield a reduced classification running time.

The uniform LBP taps the bitwise transitions existing between circular binary neighborhood to select uniform patterns from the basic LBP. Based on the belief that more can be

exploited from the circular binary pattern than just a count the number of bitwise transitions, a new feature extraction technique that establishes pairwise comparisons between consecutive digits in the circular binary pattern resulting to first-order or higher-order CD-LBP patterns is then proposed.

First-order CD–LBP– The proposed descriptor first considers the basic LBP as the zero-order CD–LBP denoted by $CD - LBP^0_{P,R}$, because the circular binary pattern is generated using a thresholding process involving circular (P, R) neighborhood pixels and the center pixel value. A new circular binary pattern is then derived as the first-order CD–LBP, $CD - LBP^1_{P,R}$, using a bitwise operation described in Figure 3.5 for a circular $(8, 1)$ neighborhood. The binary pattern $P_0 = b_7b_6b_5b_4b_3b_2b_1b_0$ is circularly shifted to

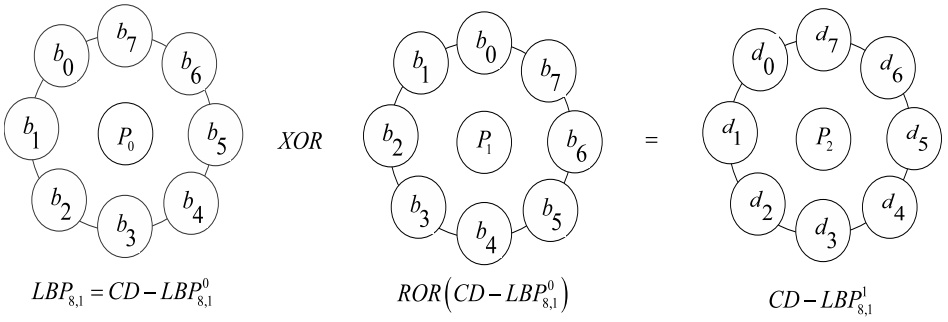


FIGURE 3.5: First-order Circular Derivative Local Binary Pattern

the right to give the binary pattern $P_1 = b_0b_7b_6b_5b_4b_3b_2b_1$, then P_0 and P_1 are XORed to produce the first-order local binary pattern $P_2 = d_7d_6d_5d_4d_3d_2d_1d_0$ corresponding to $CD - LBP^1_{8,1}$. By repeating the same operation on the binary pattern P_2 will produce the second-order CB–LBP. Then, it is possible to compute higher-order CD–LBPs.

Higher-order CD–LBP– The highest order circular derivative local binary pattern denoted by $CD - LBP^{P-1}_{P,R}$ is obtained by deriving $CD - LBP^{P-2}_{P,R}$. Meaning that a total of $P - 1$ circular derivatives that are non-zeros can be computed from a circular (P, R) neighborhood using recursive circular derivation. The n-th order CD–LBP is mathematically defined by

$$CD{-}LBP^n_{P,R} = \begin{cases} LBP_{P,R} & , \quad n = 0 \\ \{CD - LBP^{n-1}_{P,R}\} XOR \{ROR(CD - LBP^{n-1}_{P,R})\} & , \quad n \neq 0 \end{cases} \tag{3.6}$$

where (P, R) are the chosen parameters of the circular neighborhood and the order n can only vary between 0 and $P - 1$. XOR and ROR stand for the eXclusive-OR and ROtate-Right bitwise operators respectively.

Moreover, the proposed descriptor mimics the human visual interpretation in the sense that interpretability is subject to the amount of information contained in the scene or image. A grayscale image and its corresponding n-th order CD–LBP image for $n \in \{0, 1, 2, 3, 4, 5, 6, 7\}$ are depicted in Figure 3.6. As the order of the CD-LBP descrip-

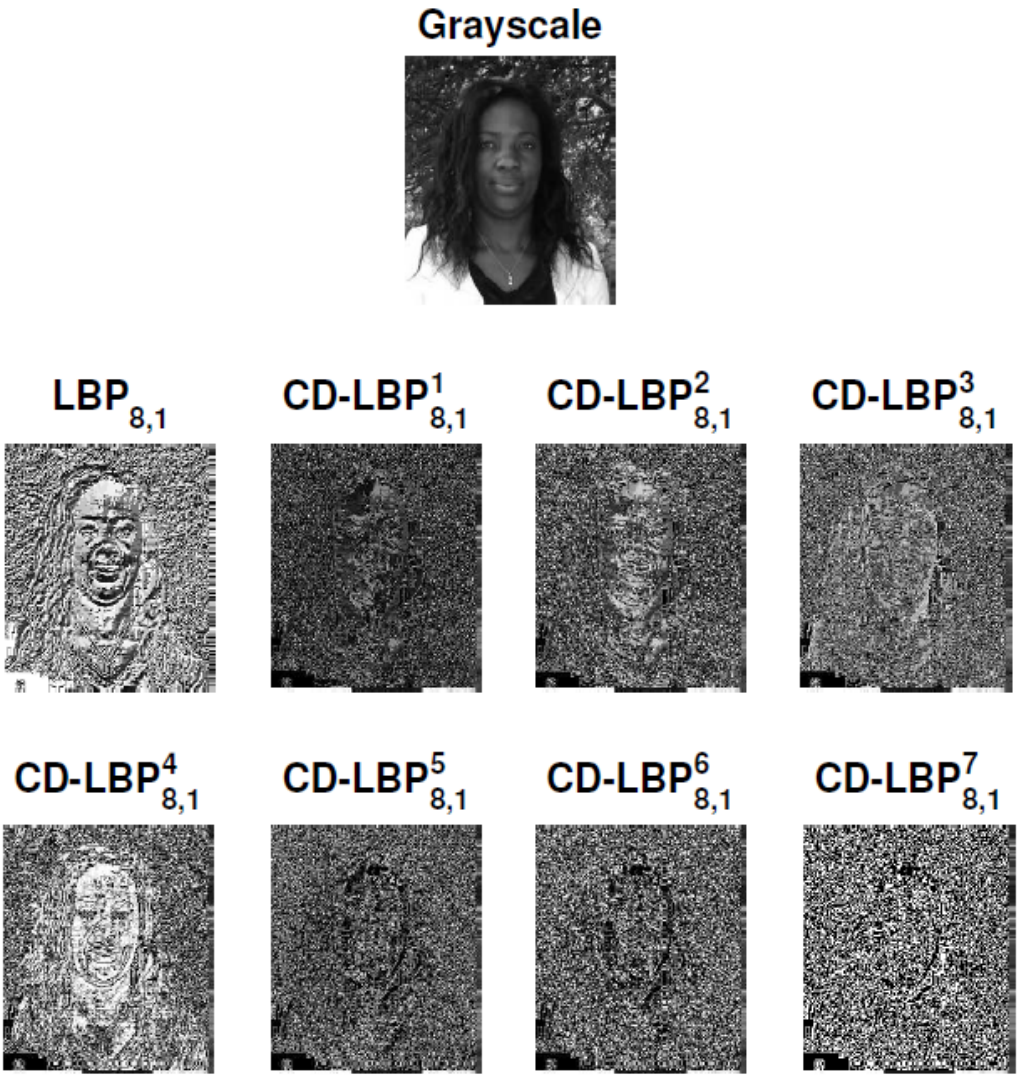


FIGURE 3.6: Example of a grayscale image and its corresponding CD–LBP images

tor increases, the information within the scene apparently lessens. The total number of possible patterns generated by n-th order CD–LBP descriptors when considering a circular neighborhood with eight sampling points is shown in Table 3.6. The patterns with respect to each descriptor, through a histogramming process as a feature representation technique, are combined into a set of features commonly called a feature vector. Although feature representation is often assumed in feature extraction, it plays an important role in providing a unique representation for every image based on the extracted features that are suitable for recognition purpose.

TABLE 3.6: Total number of patterns for each n-th order CD–LBP descriptor

	n^{th}-CD–LBP Descriptor							
Order n	0	1	2	3	4	5	6	7
Total number of possible patterns	256	128	64	32	16	8	4	2

CD–LBP Histogram– The CD–LBP histogram, which is an end result of the CD–LBP feature extraction technique, is extracted from the image in the same manner as it is done for the LBP histogram [150, 218, 219]. Considering a circular (P, R) neighborhood and an image divided into M blocks, the corresponding $CD-LBP^{n}_{P,R}$ histogram of the whole image is computed by extracting $CD-LBP^{n}_{P,R}$ features from each block and concatenating them into a single feature vector of size M x 2^{P-n}. With this method, the dimension of the $CD-LBP^{n}_{P,R}$ histogram is reduced by half as compared to the one of $CD-LBP^{n-1}_{P,R}$, which impacts positively on the running time value of the classification process. However, reduced CD–LBP histograms do not guarantee high accuracy rates. The next two sections demonstrate the relevance of the proposed CD–LBP algorithm.

3.3.2 Discrimination power of CD-LBP for pattern recognition: Application to facial expression recognition

Many researches related to facial expression analysis and recognition have been conducted in the field of computer vision for the past two decades [148, 150, 218–223]. Two mostly observed challenges in realizing a FER system similarly to HAR system its associate are the obtention of relevant facial features and the application of a suitable classifier to best demonstrate the discriminating power of the chosen features [215]. Previous literature focused more on feature extraction techniques for the main reason that extracted features should be high discriminative for better recognition accuracy, and very reduced in feature size for minimum running time or fast computational speed of the classifier. Two extensively used feature-based methods for FER are the geometric-based and appearance-based methods. Examples of Geometric feature-based methods are the active shape models (ASM) [224] and scale-invariant feature transform (SIFT) [225]. These two approaches, also referred to as local feature based approaches, have shown that the geometry-based methods rely on the detection of accurate feature points and ignore the changes such as skin texture, wrinkles and furrows [150] and they are very sensitive to noise and the accumulated errors during tracking produce inaccurate detection of features [215]. Appearance based holistic feature methods, which are less sensitive to noise and have the ability to encode micro-patterns present in the facial image, are used to overcome the shortcomings of the geometric feature-based methods.

The LBP as an appearance feature-based descriptor is an effective texture description operator that extracts the adjacent texture information in an image [41] and it has been well contemplated for FER previously [148, 218]. Resulting LBP-based handcrafted features were revealed to be highly discriminative when fed into SVM [219] and KNN [221] classifiers at the expense of high computational cost due to the high dimensionality of the feature vectors. Dimensionality reduction techniques such as the principle component analysis (PCA) [149, 215] and linear discriminant analysis (LDA) [226] were even used to reduce the dimension of the LBP histograms but resulted in an additional computational load. However, good recognition rates were reported [215, 224] despite lengthy feature size. Consequently a LBP-based feature descriptor lying in a low dimensional space is required in a FER framework to reduce the computational complexity.

3.3.2.1 Proposed method

The LBP-based FER description [148–150] is usually done after face detection in three steps: (1) the facial image is divided into various non-overlapping image blocks, (2) a LBP histogram is extracted from each block, and (3) all block LBP histograms are concatenated into a single vector used as feature vector representing the facial image. This description results in highly discriminative LBP histograms due to the considered neighborhood sampling points and the high number of divided blocks. We therefore propose a framework shown in Figure 3.7 that includes the novel CD–LBP descriptor as feature extraction technique to reduce the feature size that in turn lessens the computational burden of the classification. Two other processes accompany the CD–LBP description, which are the face detection using the Viola and Jones face detector [227] in the pre-processing stage and the use of the KNN classifier in classification stage.

3.3.2.2 Experimental results and discussions

An automatic FER experiment was carried out to demonstrate the relevance of the CD–LBP feature extraction technique. JAFFE database [228] contains 213 images of 7 facial expressions posed by 10 Japanese female models. This experiment was implemented using MATLAB and structured in three stages namely preprocessing, feature extraction and classification.

In the preprocessing stage, unnecessary features such clothes and hairstyle were removed to allow good face representation. Then facial images of 120×120 pixels were cropped from the 256×256 original images using the Viola and Jones face detector [227]. Figure 3.8 shows examples of original face images for different facial expressions (happy, anger, sad, surprise, disgust, fear, neutral) and the respective cropped images.

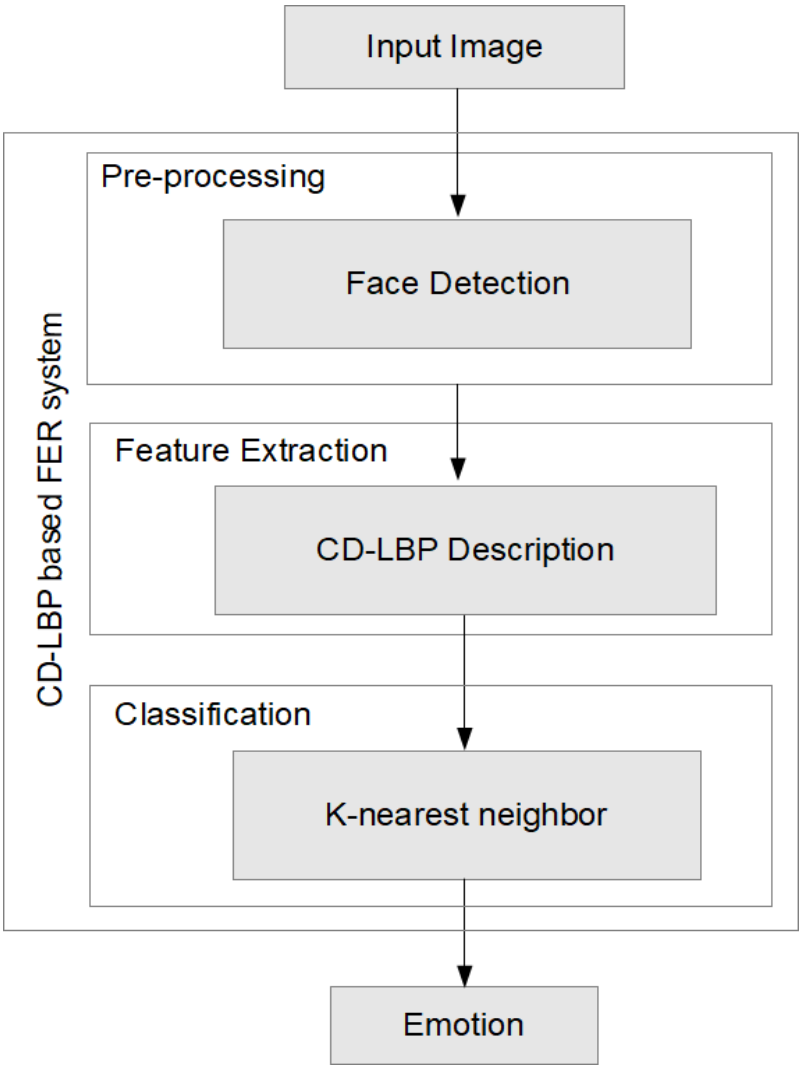


FIGURE 3.7: Proposed FER framework.

In the feature extraction stage, each cropped image was divided into 64 (8×8) blocks made off of 15×15 pixels each. Using the circular $(8, 1)$ neighborhood, CD–LBP features were extracted from each block and concatenated into a single CD–LBP histogram. This was done for orders varying from $n = 0$ to $n = 7$. A total of eight feature vectors was created in the process. The dimensions of these feature vectors are respectively 16384, 8192, 4096, 2048, 1024, 512, 256 and 128 for each facial image. The quantity of information contained in the facial images decreased as the order of the CD–LBP increased. Examples of different CD–LBP histograms of one face image are shown in Figure 3.9. A visual effect of the proposed feature extraction technique is illustrated in Figure 3.6.

Nearest neighbor (1-NN) classifier was used to evaluate the recognition accuracy achieved and running time incurred by the proposed method. Each dataset was split into training and testing sets containing 70% and 30% of 213 CD–LBP feature vectors respectively.

FIGURE 3.8: Sample images for different facial expressions and their corresponding cropped images.

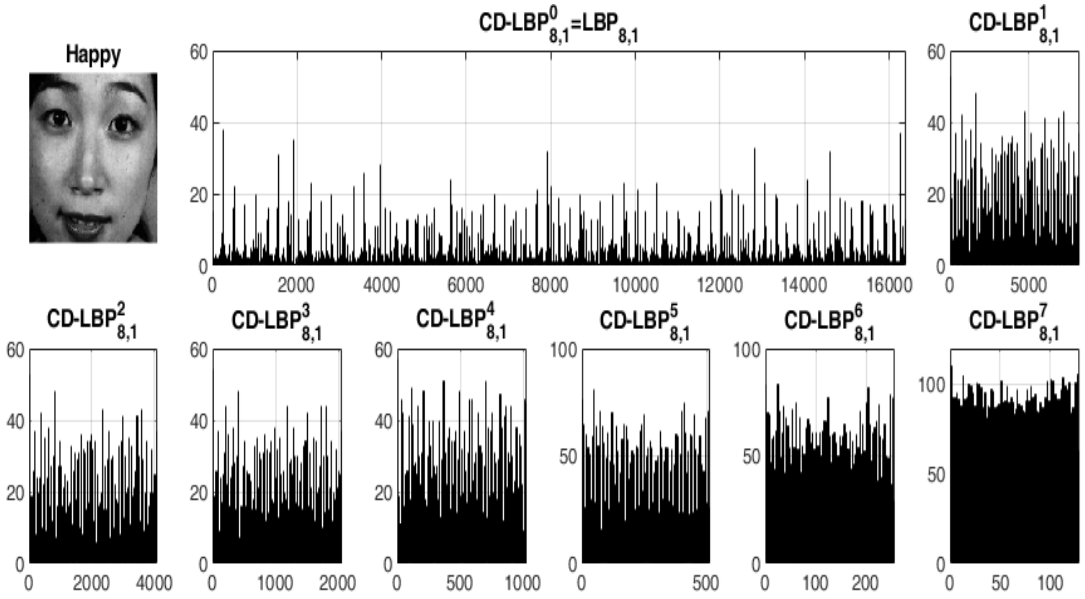


FIGURE 3.9: An cropped image and its corresponding CD–LBP feature types

The performances are shown in Table 3.7. A 100% recognition rate was obtained across all the CD–LBPs for validation. The average recognition rate of the $LBP_{8,1}$ or $CD-LBP^0_{8,1}$ is comparable to $CD-LBP^2_{8,1}$, but with a well reduced feature length and testing running time which are four times less than the one of the $LBP_{8,1}$. The $CD-LBP^3_{8,1}$ and $CD-LBP^5_{8,1}$ recognition accuracies; 89.065% and 87.500% respectively, can be acceptable, especially if the running time is of concern. A poor recognition rate was achieved by the $CD-LBP^7_{8,1}$ features, which was supported by the presence of less information.

Furthermore, the training running times do not vary much with the feature size as

TABLE 3.7: Performance results: Feature size, Mean accuracy and Running time

	n^{th}-CD–LBP Features							
Order n	0	1	2	3	4	5	6	7
Feature Size	16,384	8,192	4,096	2,048	1,024	512	256	128
Training Accuracy (%)	100	100	100	100	100	100	100	100
Training RunTime (sec)	0.4416	0.4351	0.4292	0.4420	0.4452	0.4479	0.4361	0.4345
Testing Accuracy (%)	90.625	89.065	90.625	89.065	81.250	87.500	78.125	28.125
Testing Run Time (sec)	0.1960	0.1076	0.0565	0.0316	0.0199	0.0139	0.0088	0.0086

compared to the testing running times whereas a considerable decrease in running time is observed. The steadiness of the training running time is justified by the learning process of the classifier. Figure 3.10 illustrates the running time variation for both training and testing.

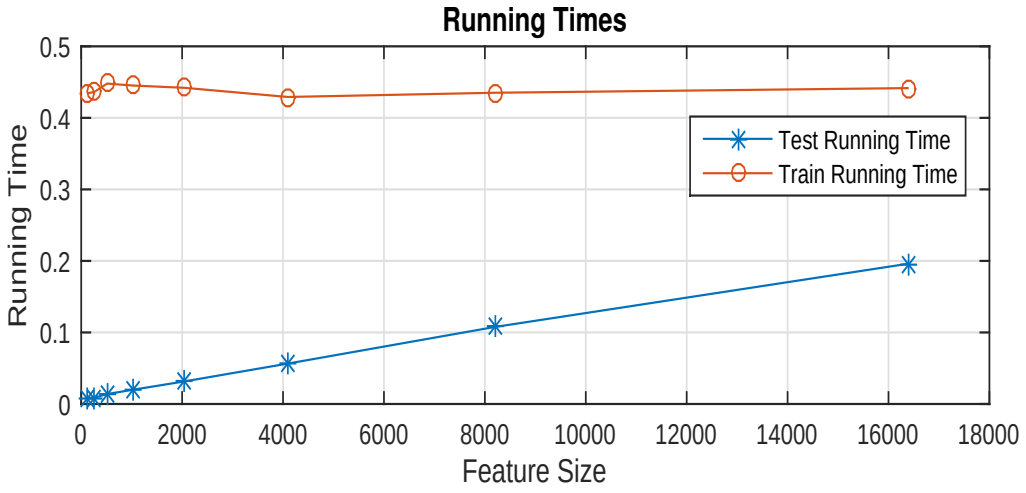


FIGURE 3.10: Relation between Feature Size and Running Time

Our results are compared to the ones of three state-of-the-art methods applied to the JAFFE dataset. In [229], an Active Appearance Model (AAM) was used to locate the facial point automatically. The relevant facial images were processed then LBP was used to extract features that were fed to a KNN classifier. An recognition rate of 72% was obtained with feature vector dimension of $19,456$ derived from the concatenation of histograms for 76 non-overlapping image blocks. This recognition rate is lower than the one obtained using the 6^{th}-order CD-LBP features of dimension 256 which is far less than $19,456$. In [230], feature descriptors from three variants of LBP; namely MLBP, LBPV and CS-LBP, were extracted and fed to a KNN classifier; recognition rate achieved were compared to the one achieved when LBP features. The CS-LBP provided the best recognition rate of 87% with an estimated feature vector dimension of $14,400$. Though the CS-LBP feature descriptors are of low dimension than the LBP, its applications here provided a high dimension due to the number of divided blocks in

an image. The recognition rate achieved by the CS-LBP features is close to the one obtained with the 5^{th}-order CD-LBP features of a dimension 512. In [231], the authors first applied the LBP-KNN scheme to the JAFFE database which led to a recognition rate of 83.66%, then later proposed an abstract extraction of facial features from LBP features through an Deep Belief Networks (DBNs) followed by the classification with the Softmax classifier. Though their proposition increased the recognition rate by about 4%. The use of DBNs required an additional computational effort. Summing up, the use of the proposed method outperforms the results obtained in the selected state-of-the-art works in terms of both recognition rate and computation speed.

3.3.3 Application of CD-LBP to pedestrian action recognition

Researches on pedestrian action recognition (PAR) have been actively carried out in the last decade; the main aim being to reduce cases of traffic related injuries and fatalities, specially with the emergence of various intelligent driving assistance systems [16, 107, 210, 211, 232, 233]. As vehicles need to interpret pedestrian actions before making appropriate decisions in response, a PAR system which is robust can then ensure cautious and stable operations of intelligent driving systems in complex or uncertain environments. Such system requires not just the detection of pedestrians, but also the recognition of their behaviors or actions for an informed decision. These two processes; detection and recognition, are comparable to those used in HAR. The recognition of pedestrian actions therefore stems from the recognition of human actions and inherits HAR related challenges discussed in Chapter 2. Moreover, the detection of pedestrian action, corresponding to action representation in a HAR scheme, has been based on handling of spatio-temporal templates with various feature descriptors.

In [16], pedestrian activities during road crossing are classified using patterns of motion and HOG descriptor. The MHI templates as primary holistic representation of video sequence were used and handcrafted features extracted from them using HOG descriptor fed into a SVM classifier. Although satisfactory recognition rate of 91.4% was achieved for the MHI-HOG-SVM model applied to the pedestrian action dataset presented in Section 2.6, this rate was outperformed in a proposed model symboled MHI-LBP-HOG-SVM [107] with a 94.3% as recognition rate. This improvement highlighted the importance of LBP descriptor which, besides its computational simplicity, was solicited because of its ability to encode the direction of motion from non-monotonous regions of MHI templates. However, due to the need of a viable PAR system, we propose the use of CD-LBP descriptor and explore its potential benefit in the recognition of pedestrian actions.

3.3.3.1 Proposed method

The method proposed for PAR, whose framework is shown in Figure 3.11, is structurally

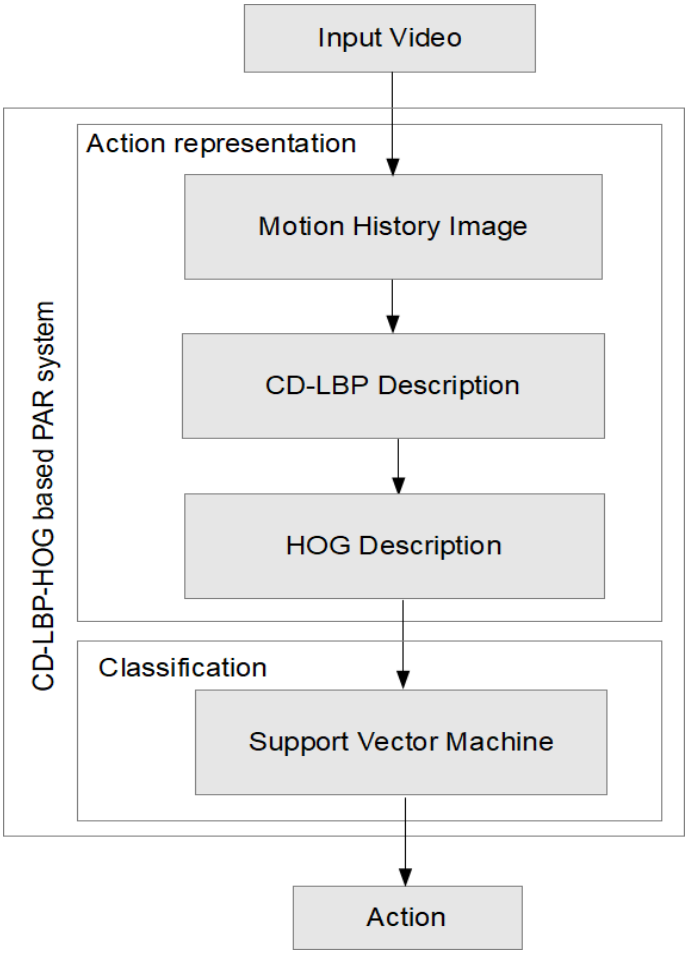


FIGURE 3.11: Proposed PAR framework.

similar to the one used in [107] presented in two stages. The first is an action representation stage: it consists in reducing a video sequence into a holistic representation in the form of a MHI image; in transforming the MHI image into a LBP image, then in extracting handcrafted features from the latter using HOG descriptor. The second is a learning stage which consists of a SVM classifier used for the recognition of pedestrian actions. The particularity of our proposition is the substitution of the LBP image by CD–LBP image produced by our proposed CD–LBP algorithm.

Action representation– As mentioned in Chapter 1, a HAR system or even PAR system owes its best performance to a proper representation of the action. So in the proposed method, the MHI representation which presents some advantage contrarily to motion detection methods [55, 83], participates in describing an action. As opposed to motion detection method, the generated temporal template from the MHI principle summarizes motion information of several frames in a spectrum of values not just limited

to binary values. The aggregation of motion information from preceding video frames contributes to the robustness of the PAR system when a pedestrian is moving for a short duration. The novel CD–LBP algorithm is then applied to MHI templates to deal with appearance-based problems such as illumination variation, clothes change, background clutter, etc., while producing CD–LBP images denoted by $CD - LBP_8^n$ with n varying from 0 to 7 being the order of the CD–LBP feature descriptor. In Figure 3.12, a MHI

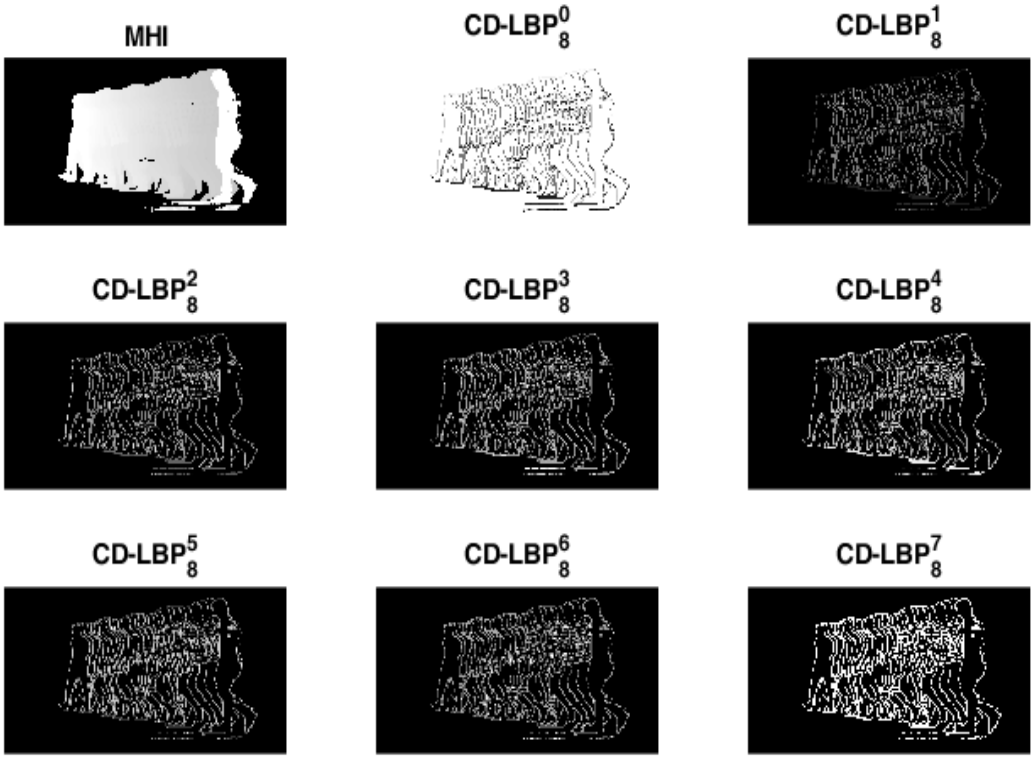


FIGURE 3.12: A cross walk MHI image and its corresponding CD–LBP images.

image and its corresponding n-order CD–LBP images highlighting the sharpening effect on the edges for an cross walk action are depicted.

After the CD–LBP creation, handcrafted features are then extracted using the HOG descriptor which is widely used to detect object in still images. This descriptor was initially proposed by Dalal and Triggs [55] and later employed in human detection algorithms [234]. With its ability to describe the local area within the image in terms of object appearance and shape, and the distribution of the edge orientation in terms of intensity gradient, the HOG descriptor counts occurrences of gradient orientation. Following the description of the HOG-based feature extraction technique discussed in Chapter 2, CD-LBP images are resized to $P \times Q$ pixel images which are then divided into cells of size $N \times M$ pixels. The magnitude and orientation at each pixel location within each cell are computed using Equations 2.14 and 2.15. These two computed quantities are later used in the accumulation of orientations of all pixels to form an K-bins histogram per

cell ranging from 0 to 180 degrees. In order to construct a suitable action representation for recognition, all cell histograms are therefore concatenated into a unique histogram containing $\frac{P}{N} \times \frac{Q}{M} \times K$-handcrafted features. The HOG feature extraction process is illustrated in Figure 3.13. However, as the size of the feature descriptors relates to the

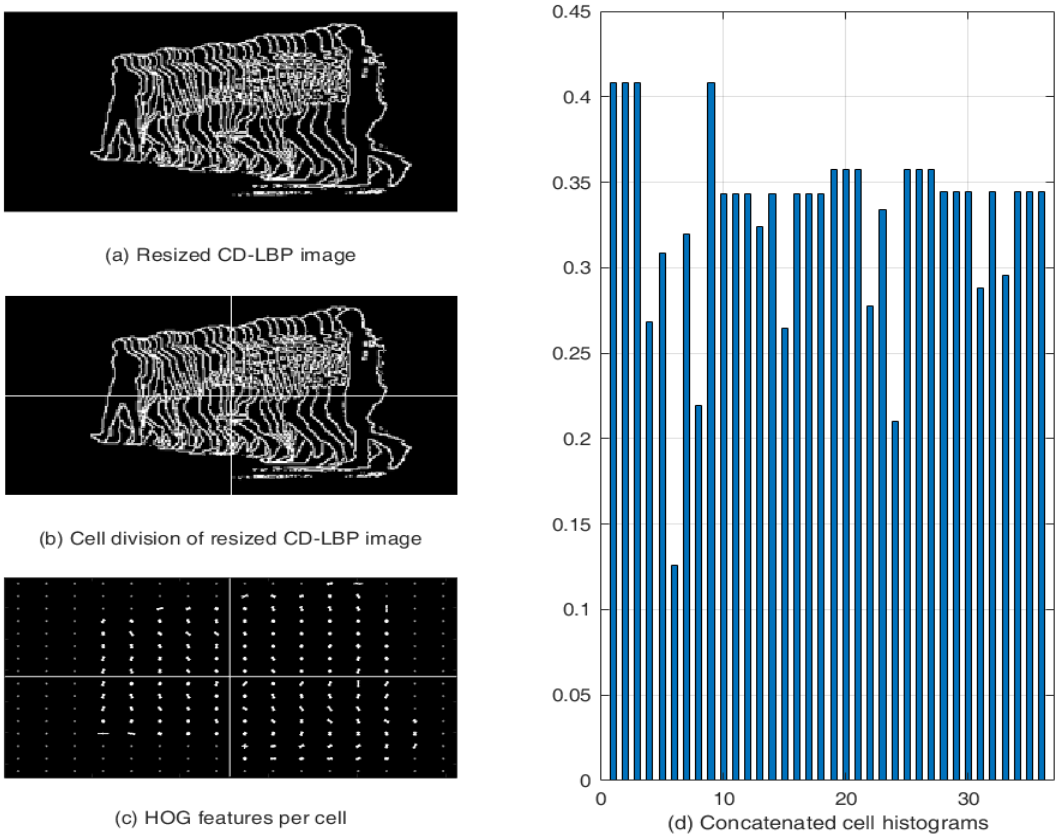


FIGURE 3.13: HOG feature extraction process: (a) a resized CD–LBP image, (b) 2×2 block consisting each of one cell, (c) HOG features per cell, and (d) the histogram of oriented gradients corresponding to the concatenation of all four 9-bins cell histograms.

computational complexity of the PAR system, the parameters P, Q, N, M and K are therefore chosen experimentally in relation to a improved state-of-the-art performance.

Classification– For recognition purpose, the SVM is used. This classifier is one of state-of-the-art large margin classifiers recently exploited in visual pattern recognition applications [17], which also include the pedestrian activity recognition [107]. Linear SVM classifiers are then used to train and test various derived CD–LBP-based HOG features. As SVM is a binary classifier, a multi-class classification approach is then implemented. A simple and less computational expensive one-against-all scheme compared to other cross-validation schemes is considered to analyze performance of n-order CD–LBP-based HOG feature sets constructed as pedestrian action representations.

3.3.3.2 Experimental results and discussions

In this section, the proposed PAR framework in Figure 3.11 is implemented and tested using Pedestrian action database presented in Section 2.6.2 to demonstrate the relevance of our proposed CD-LBP algorithm. The database contains a total of 160 video clips scaled at a resolution of 170×320 of 20 persons performing each 8 different actions. The effectiveness of the proposed method is then evaluated by carrying out an experiment using the MATLAB scientific programming language. This experiment is conducted in two stages namely action representation and classification. It is also repeated a number of times with respect to; (1) the order of the CD–LBP descriptor, (2) the number of divided cells of CD–LBP images during HOG feature extraction, and (3) the number of action classes considered during classification.

Results on action representation– From each video sequence, a total of 30 frames are processed to generate a MHI image. The values of the key parameters, $\tau = 255$, $\epsilon = 30$ and $\delta = 2$, that have yielded a computation of MHI images and capturing of relevant motion information were selected to reduce the effect of MHI related issues [6]. Moreover, as the MHI method is an appearance-based method, the proposed CD–LBP alogrithm is then applied to MHI images to produce CD–LBP images of size 168×318, which in turn are used to extract HOG features. Figure 3.12 shows a MHI sample image and its corresponding $CD - LBP_8^n$ images for a circular $(8, 1)$ neighborhood and values of n ranging from 0 to 7. When observing comparing the CD–LBP images at different order, the quantity of information decreases with respect to the increase of the order. Using the entropy of an image, defined as

$$H = -\sum_{k=1}^{K} p_k log_2(p_k), \tag{3.7}$$

where K is the number of gray levels and p_k is the probability associated with the presence of gray level k, we measure the amount of information of the $CD-LBP_8^n$ images in Figure 3.12. The measured entropy results are depicted in Figure 3.14. Although $CD - LBP_8^7$ image corresponds to the image with the lowest entropy, it appears to be the most visual and descriptive image of all $CD - LBP_8^n$ images for $n \in [0, 7]$ because of the brightness observed in all edges within the image. Despite self-occlusions that occur while performing the actions, shapes described by those actions are preserved in the $CD - LBP_8^n$ images for $n \in [0, 7]$. This preservation is then exploited in the extraction of HOG handcrafted features as action representation. We consider four cell sizes in our search of the most discriminative HOG descriptor. The properties of each HOG descriptor with respect to the cell division are tabulated in Table 3.8. The non-overlapping of blocks (**BlockOverlap** $= 0$), while getting the 9-bins histogram per cell

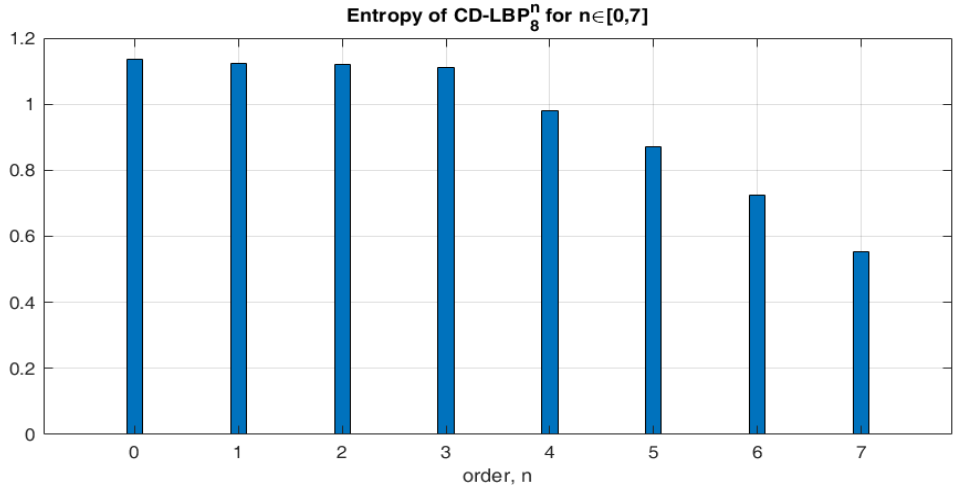


FIGURE 3.14: Entropy of Cross Walk CD–LBP images shown in Figure 3.12.

TABLE 3.8: Properties of HOG descriptors with respect to $D \times D$ cell division.

	$D \times D$ **HOG Descriptors**			
HOG description ($D \times D$ cells)	2×2	4×4	8×8	16×16
CD-LBP image ($P \times Q$ pixels)	128×256	128×256	128×256	128×256
Cell/ Block size ($N \times M$ pixels)	64×128	32×64	16×32	8×16
BlockOverlap	0	0	0	0
Total number of cells/ blocks	4	16	64	256
Number of bins (K)	9	9	9	9
HOG feature size ($\frac{P}{N} \times \frac{Q}{M} \times K$)	36	144	576	2304

or block, contributes to a reduced size of the concatenated HOG histogram. A total of 32 feature sets are created after we have applied the four HOG descriptors to the eight sets of images produced by the CD–LBP algorithm. Figure 3.15 displays how descriptive is the HOG features visually with respect to the number of cells. Followed

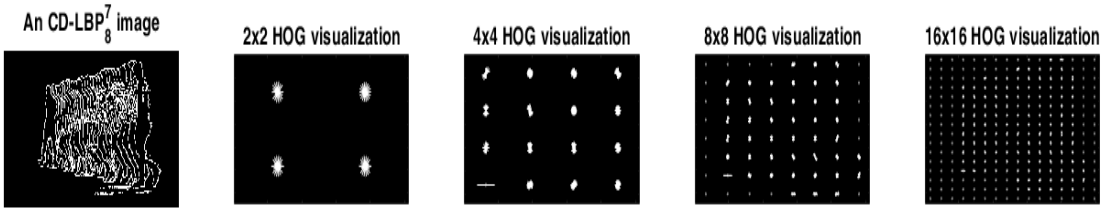


FIGURE 3.15: Visualization of $D \times D$ HOG descriptors for $D = 2, 4, 8, 16$.

by the 8×8 HOG features, the 16×16 HOG features extracted from the $CD - LBP_8^7$ image set delineate the shapes caused by the action performed. They also constitute the most descriptive HOG descriptor despite their lengthy feature size (See Figure 3.16). However, the discriminative power of each feature set is tested using the SVM classifier.

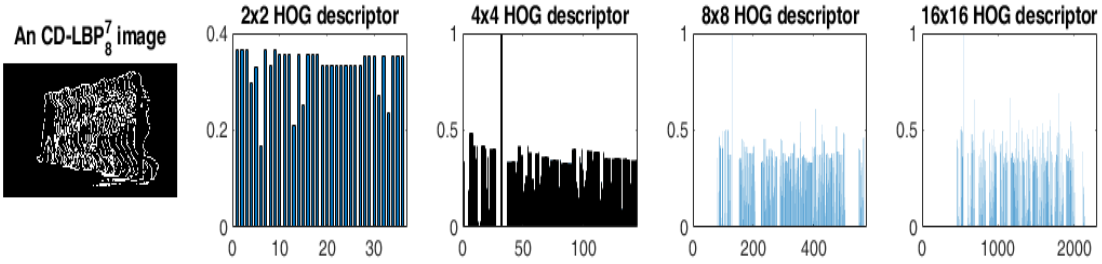


FIGURE 3.16: Examples of histogram for $D \times D$ HOG descriptors, for $D = 2, 4, 8, 16$.

Results on classification– Linear SVM classifiers in conjunction with HOG features are used here to evaluate the classification performance in terms of recognition rate. As each feature set denoted by $D \times D - CDLBP_8^n - HOG$ for $D = 2, 4, 8, 16$ and $n = 0, 1, 2, 3, 4, 5, 6, 7$ is fed into a one-against-all SVM multi-class classifier for learning purpose, a total of 32 classifications are then performed. The resulting accuracies for 8 action classification are recorded in Table 3.9 and the highest accuracies with respect to each $D \times D - CDLBP_8^n - HOG$ feature set are highlighted in bold. A visual rep-

TABLE 3.9: Classification accuracies with respect to $D \times D - CDLBP_8^n - HOG$ for $D = 2, 4, 8, 16$ and $n = 0, 1, 2, 3, 4, 5, 6, 7$.

Parameters $D \times D$ **and** n	0	1	2	3	4	5	6	7
2×2	48.12	45.00	52.50	51.87	53.12	50.00	46.25	**60.62**
4×4	63.12	72.50	69.37	68.12	71.87	73.12	68.75	**77.50**
8×8	73.12	79.37	76.87	77.50	82.50	78.12	78.75	**82.50**
16×16	86.87	90.00	91.25	93.12	89.37	90.00	91.87	**94.37**

resentation of resulting accuracies is also provided in Figure 3.17 for accuracy analysis. Classification accuracies increase consistently with the number of $D \times D$ cell division for each $CD - LBP_8^n$ images for $n = 0, 1, 2, 3, 4, 5, 6, 7$. This is not the case when comparing classification accuracies with respect to the n-order CD–LBP images for each considered $D \times D$ HOG descriptor for $D = 2, 4, 8, 16$. This inconsistence is particularly observed for HOG features derived from $CD - LBP_8^n$ images for $n = 1, 2, 3, 4, 5, 6$ and may be justified by the level of complexity of those feature sets. However, the resulting accuracies from the $D \times D - CDLBP_8^7 - HOG$ feature sets for $D = 2, 4, 8, 16$ are higher than those obtained from $D \times D - CDLBP_8^n - HOG$ feature sets for $D = 2, 4, 8, 16$ and $n = 0, 1, 2, 3, 4, 5, 6$. And among these high accuracies, the highest with a score of 95% is achieved by the $16 \times 16 - CDLBP_8^7 - HOG$ feature set which in turn has a discriminative power higher than the other features.

In addition, the confusion matrix corresponding to the highest recognition rate is shown in Table 3.10. This matrix highlights a very low rate of false recognition in run, walk, fall down, fall and rise up, and turn towards camera actions, which may be due to

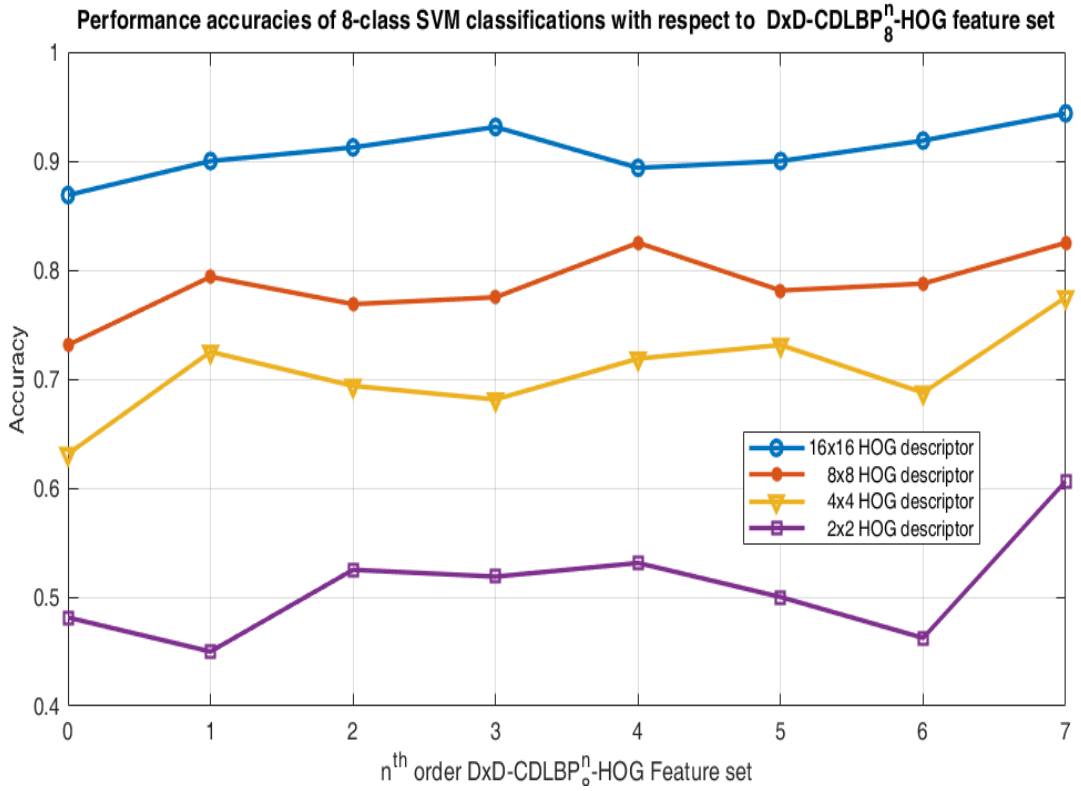


FIGURE 3.17: All 8-class SVM classification performances with respect to $D \times D$ HOG descriptors, for $D = 2, 4, 8, 16$.

some similarities existing between their respective features. Besides occlusion being a MHI limitation and also occurring in actions such as cross walk, halfway return, and turn opposite to camera, a 100% recognition rate is obtained. Consequently, the average of all

TABLE 3.10: Confusion matrix for the 8-class classification with the highest average recognition rate of 94.37%.

Actions	CrsWlk	FalUp	FalDw	HafRt	Run	TrnOp	TrnTo	Walk
CrsWlk	100	0	0	0	0	0	0	0
FalUp	0	95	0	0	0	0	0	5
FalDw	0	0	95	0	0	5	0	0
HafRt	0	0	0	100	0	0	0	0
Run	0	0	0	0	85	5	0	10
TrnOp	0	0	0	0	0	100	0	0
TrnTo	0	0	0	0	0	5	95	0
Walk	0	5	0	0	10	0	0	85

recognition rates corresponding to the 8 actions is still high and our proposed method that we symbolize by MHI-CDLBP-HOG-SVM outperforms state-of-the-art methods applied on the Pedestrian Action dataset. Table 3.11 compares our proposed method to other previous different methods. Results from MHI-HOG-KNN and MHI-HOG-SVM

TABLE 3.11: Comparative performance accuracies on Pedestrian Action dataset

Method	Recognition accuracy (%)
MHI-HOG-KNN [16]	74
MHI-HOG-SVM [16]	91
MHI-CDLBP-HOG-SVM	**94.37**

methods once again confirm the suitability of HOG features for the SVM classifier. This may be because of the ability of SVM to learn from high dimensional features.

However, the idea to use the CD–LBP algorithm comes from the work of Ahad et al. [107] who proposed a MHI-LBP-HOG-SVM method to improve the accuracy of pedestrian actions. But their improvement of the recognition rate is achieved by considering only 7 actions as the fall and rise up action is discarded. For this reason, our experiment is then repeated for 7 actions. Table 3.12 shows the confusion matrix of the best recognition rate of all the thirty-two 7-class SVM classification recognition rates depicted in Figure 3.18.

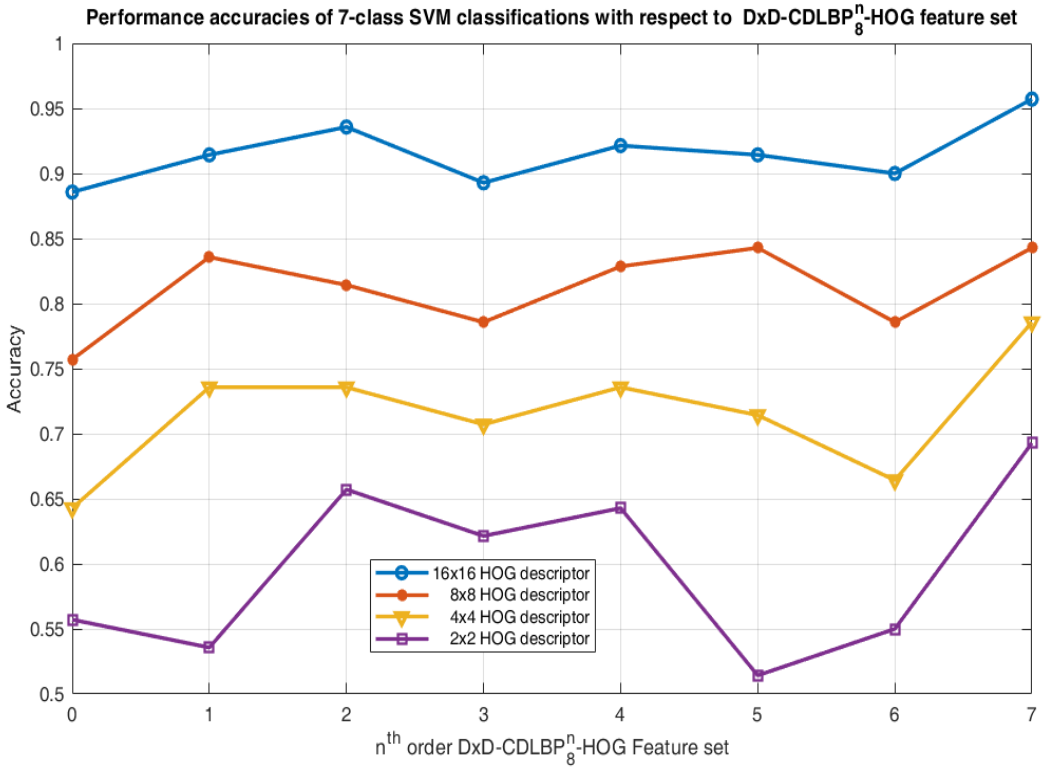


FIGURE 3.18: All 7-class SVM classification performances with respect to $D \times D$ HOG descriptors, for $D = 2, 4, 8, 16$.

When analyzing the thirty-two 7-class SVM classifications, the resulting recognition rates present a similar trend to those of the thirty-two 8-class SVM classifications in terms of $D \times D - CDLBP_8^n - HOG$ for $D = 2, 4, 8, 16$ and $n = 0, 1, 2, 3, 4, 5, 6, 7$. The

TABLE 3.12: Confusion matrix for the 7-class classification with highest average recognition rate of 95.71%.

Actions	CrsWlk	FalDw	HafRt	Run	TrnOp	TrnTo	Walk
CrsWlk	100	0	0	0	0	0	0
FalDw	0	100	0	0	0	0	0
HafRt	0	0	100	0	0	0	0
Run	0	0	0	85	5	0	10
TrnOp	0	0	0	0	100	0	0
TrnTo	0	0	0	0	5	95	0
Walk	0	0	0	10	0	0	90

recognition rates still grow proportionally to the number of divided cells and the best recognition rate is reached for the $16 \times 16 - CDLBP_8^7 - HOG$ feature set. Although the reason for not including the fall and rise up action is not mentioned in the literature, our proposed method with a recognition rate of 95.71% still outperforms various improved methods employed in [107]. See Table 3.13 for comparison. In addition, the robustness

TABLE 3.13: Comparative performance accuracies on Pedestrian Action dataset

Method	Recognition accuracy (%)
MHI-HOG-KNN [107]	78.66
MHI-HOG-SVM [107]	91.42
MHI-LBP-HOG-SVM [107]	94.30
MHI-CDLBP-HOG-SVM	**95.71**

of $CD - LBP_8^7$ feature set is proven in the experiments conducted in this work. Besides the fact that the $CD - LBP_8^7$ descriptor in conjunction with HOG descriptor demonstrates a high discriminative power yielding high recognition rates when compared to the original LBP descriptor, its binary representation can be beneficial in terms of memory complexity in a real-time implementation.

3.4 Summary

In this chapter, we proposed some handcrafted feature based approaches as contributions for the recognition of human actions. These methods, apart from addressing issues such as illumination variations, scale changes, viewpoint variations, and background clutter, aimed at reducing the computational complexity experienced in HAR systems. So in Section 3.2, the SPS algorithm [43] was employed in a simple framework to recognize human action. This algorithm, which is a holistic feature extraction, took advantage of the presence of slope patterns in MHI templates to extract them as relevant features for action representation. These features, also referred to as SPS handcrafted features of reduced size, were fed into a KNN classifier for recognition purpose. The KTH Action

dataset was used to validate the proposed method and the per scenario classification results were quite encouraging and showed the relevance of the SPS feature as holistic action representation. Then in Section 3.3, we proposed a novel descriptor called circular derivative local binary pattern (CD–LBP), which has the ability to provide lower dimension and high discriminative feature vectors as handcrafted features. This descriptor is developed by evaluating the relation between consecutive binary digits representing the circular binary structure . The relevance of CD–LBP features was then demonstrated in two applications. In the first application (FER), the CD–LBP algorithm was applied to the JAFFE dataset to generate higher-order CD–LBP features which in turn were fed into the KNN classifier. As for the second application (PAR), video sequences of pedestrian actions were considered to capture motion information in the form of MHI templates. The CD–LBP in conjunction with the HOG descriptor were applied to MHI images to extract features which were later used to learn a one-against-all linear SVM classifier. In both applications, improved performance were therefore achieved.

Chapter 4

Contribution to Deep learning feature based approach to human action recognition

'Thought and learning are of small value unless translated into action"

Wang Yangming

4.1 Introduction

Nowadays, deep neural networks are widely used for human action recognition (HAR) due to their ability to operate directly on the raw video inputs by extracting both the spatial and temporal information. Contrary to traditional approaches which are highly problem-dependent and constrained in real-world applications, the deep learning based techniques build a high level of representation directly from the query video input without any pre-processing step. A popular deep learning model is the CNN-based model which relies on the adaptation of multilayered neural deep architectures to process real-world data. This model comes in various dimensions depending on the data at hand. The 2D-CNN model handles 2D raw inputs corresponding to video frames. This type of CNN model can only learn the spatial information and ignore the motion information encoded in multiple adjacent video frames. Furthermore, 3D-CNN models are used to incorporate the motion information by capturing both spatial (2D) and temporal (1D) dimensions composing videos. However, 3D-CNN models that have achieved superior performance for video classification are computational costly, making it necessary to

seek for simplified deep neural network models for HAR, especially in absence of adequate computational resources such as fast central processor units (CPUs) or graphical processor units (GPUs). Therefore, as a contribution in the field of deep learning, we propose a method that reduces computational complexity imposed by the 3D-CNN approaches while addressing the problem of lack of motion information when applying the 2D-CNN model to video frames. This method recognizes human actions by using a simple 2D-CNN model that learns robust feature representation from temporal information embedded into the motion history images of action videos. The KTH action database is used to validate our proposition and the achieved results compared favorably against the hand-crafted state-of-the-art methods.

The rest of this chapter is organized as follows. The proposed deep learning based approach is presented in Section 4.2 followed by experimental results and discussions in Section 4.3. Finally the chapter is summarized in Section 4.4.

4.2 Proposed method

In this section, the proposed method and its related components are presented. This method reduces the computational complexity highlighted in Chapter 2 when applying the deep learning feature based approaches to HAR. This reduction is done by introducing a pre-processing step into a 2D-CNN architecture. The proposed approach comprises two stages:

- The pre-processing stage which captures motion information from video in the form motion history image (MHI) plays an important role. It reduces the number of 2D-CNN operations performed on each video frame or prevents the use of 3D kernels in a 3D-CNN based approach.
- The recognition stage is where hidden behavioral features are learned and classified. The flowchart illustrating the proposed method is depicted in Figure 4.1.

4.2.1 Pre-processing stage

The integration of this stage in a 2D-CNN architecture justifies the innovation of our proposed deep learning approach. Video sequences are reduced to a single image known as MHI template, that keeps the history of temporal changes at each pixel location, which then decays over time [212]. This template is in grayscale value indicating the

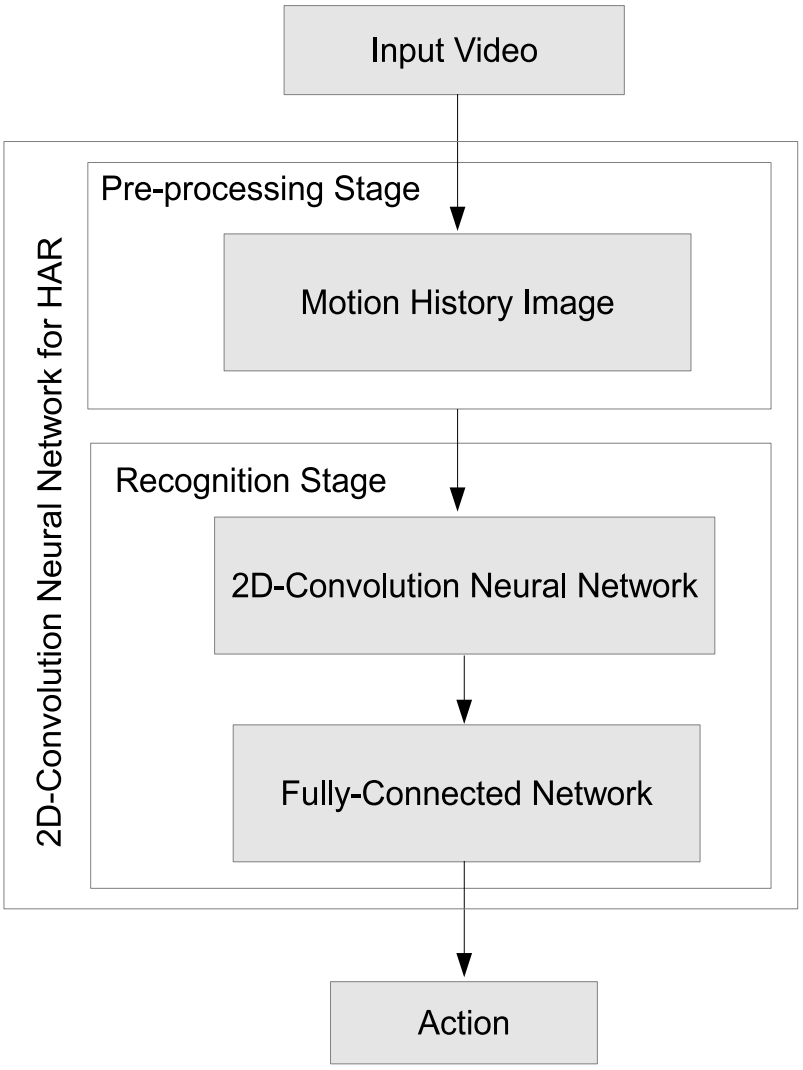


FIGURE 4.1: Proposed methodology.

most recent motion of the pixel in a set of video sequences. Recent motions are represented by brighter pixel intensity values in the MHI [213]. Equations 2.11, 2.12, and 2.13 defined earlier govern the principle behind this holistic representation.The selection of the involved parameters plays an important role in generating the MHI images for a good motion representation of actions. Although computation of MHI templates is inexpensive, there is a problem of occlusion for some human actions where motion information is overwritten . This problem is addressed here by using the 2D-CNN model, which has the ability to learn features, including hidden features from the handcrafted approach. Figure 4.2 shows an *handclapping* example of MHI.

4.2.2 Recognition stage

A 2D-convolutional neural network is used here as learning algorithm to recognize human actions from their respective MHI images. This learning algorithm, introduced by LeCun

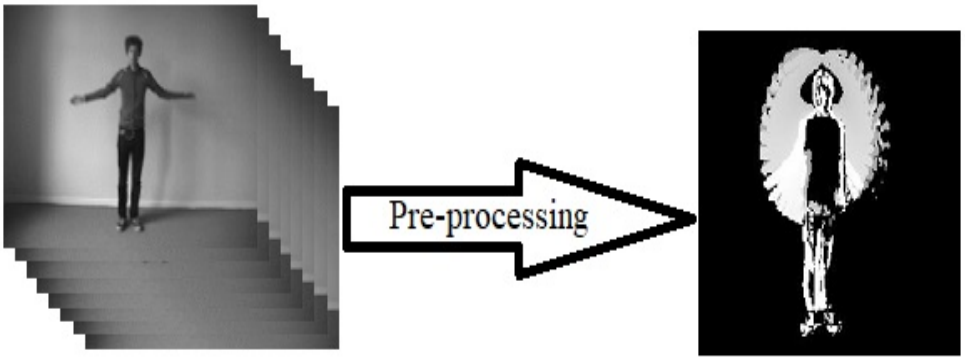


FIGURE 4.2: Extraction of the motion history image (MHI).

et al. [11], is known as a supervised learning based multistage deep neural network and performs both action representation and classification. The action representation refers to the feature extraction phase which is basically composed of the convolution and pooling layers, giving the CNN the ability to learn multiple stages of invariant features from input images. The classification phase consists of a fully connected layer with all activations in the previous convolution layer, which converts flattened feature maps into feature vectors. Lastly, the softmax layer from the classification phase outputs the action classes.

Our deep learning architecture, which is simple and comparable to the one of the popular LeNet architecture used for character recognition [11], is composed of two sets of combined convolution and pooling layers, and two full connected layers.

- **Convolution layer** – This layer takes in an image from the previous layer, which is then convolved with a set of the 2D kernels, filters or weights and produces feature maps whose the number depends on the number of 2D kernels considered. In addition, each of the 2D kernels is shared across the entire resulting feature map. The convolution operation is executed by sliding the 2D kernel over the input image, where at each location a matrix multiplication is performed and result added onto the feature map.

- **Pooling layer** – After a convolution layer, a pooling layer is usually added in between CNN layers with unique function to reduce the dimensionality of feature maps yielding to the reduction of both number of parameters and computation in the network. The max-pooling, where the maximum activation in the sub-sampling region is kept as the significant information, is used here.

- **Full connected layers** – They are the last layers of the 2D-CNN after the convolution and pooling layers, and constitute the classification phase. These layers only accept 1D data, hence the feature maps from the pooling layer are converted

into 1D feature vector. The phase of the 2D-CNN operates exactly as a regular neural network where the neurons in the full connected layer have full connections to all the activations in the preceding layer. Here, the softmax activation function is used in the very last layer (output layer) of the 2D-CNN to produce the probability of each action class given its corresponding MHI image.

The proposed 2D-CNN architecture, shown in Figure 4.3, from top to bottom, is de-

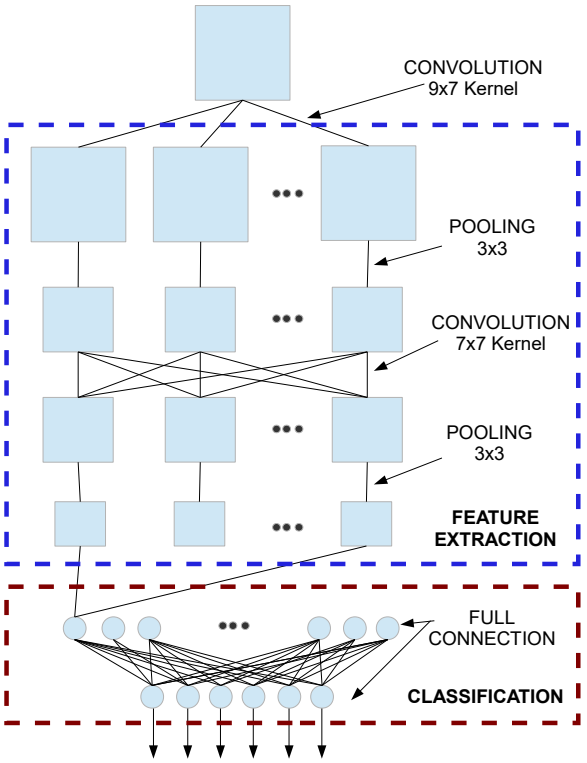


FIGURE 4.3: Proposed 2D-CNN architecture.

scribed as follows: the first layer corresponds to the input layer, which takes a MHI image of size 80×60 as input. This input is then convolved with pre-defined number of kernels of size 9×7, producing a feature map of same size as the input image for each kernel. The resulting feature maps form the second layer which is the convolution layer. These feature maps are sub-sampled by max-pooling of size 3×3, producing new feature maps with reduced individual size at the third layer, also referred to as pooling layer. A second convolution with another pre-defined number of kernels of size 7×7 is effected at the fourth layer followed by a pooling layer which again reduces the size of each feature map with a max-pooling size of 3×3. At the fifth layer, feature maps from the fourth layer are flatten and encoded in a vector of size 128. This vector can be interpreted as a feature descriptor of salient spatio-temporal information embedded in the MHI image. Finally, the last layer contains a classical multilayer perceptron with

one neuron per a specific class of action, thus a total of six neurons. This layer is also referred to as the output layer with softmax function as activation function.

4.3 Experimental results and discussions

In our experiment, we use the KTH action dataset presented in Section 2.6.3 to evaluate the proposed 2D-CNN approach for HAR. The Matlab software is used to generate a new dataset of MHI templates from the KTH action dataset. Then later the recognition stage, which entailed both the 2D-CNN (action representation) and the full-connected network (classification), is implemented under the Python programming language using Keras API with TensorFlow library as a backend. The new MHI dataset generated in the pre-processing stage using Matlab is later divided into two sets, a training set and a testing set.

4.3.1 Action representation

The 2D-CNN model is referred here to as feature extraction method. Figure 4.3 shows how features are learned from the input MHI templates which are computed using the MHI parameters $(\tau, \epsilon, \delta) = (255, 40, 2)$ while considering fifty frames per video [34]. An example of the MHI template resulting from the preprocessing stage is shown in Figure 4.2. MHI templates per scenario can also be observed in Figure 3.2 in Chapter 3. The parameters of each layer in the proposed 2D-CNN as discussed in Section 4.2 are presented in Table 4.1. A total of sixteen randomly selected kernels are used in

TABLE 4.1: Parameters for the proposed 2D-CNN structure

Layer	Type	Feature maps	Kernel size
1	Convolution	16	9x7
2	ReLU	16	none
3	Pooling	16	3x3
4	Dropout	16	none
5	Convolution	32	7x7
6	ReLU	32	none
7	Pooling	32	3x3
8	Dropout	32	none

the first convolution and thirty-two for the second convolution. We resize the MHI templates to 80×60 to reduce the memory requirement of our 2D-CNN model. The two convolutional layers use kernel sizes 9×7 and 7×7 respectively, and the two pooling layers use kernels of size 3×3. To enhance the performance of our deep learning model, we additionally introduce a ReLU (Rectified Linear Unit) layer in-between the set of

convolution and pooling layers, and a Dropout layer assuming a 25% dropout rate at the end of each pooling layer for regularization purpose. An image from feature maps for a MHI template of a *Boxing* action that illustrates each layer is shown in Figure 4.4. In addition, the output shape and the number of parameters with respect to each layer

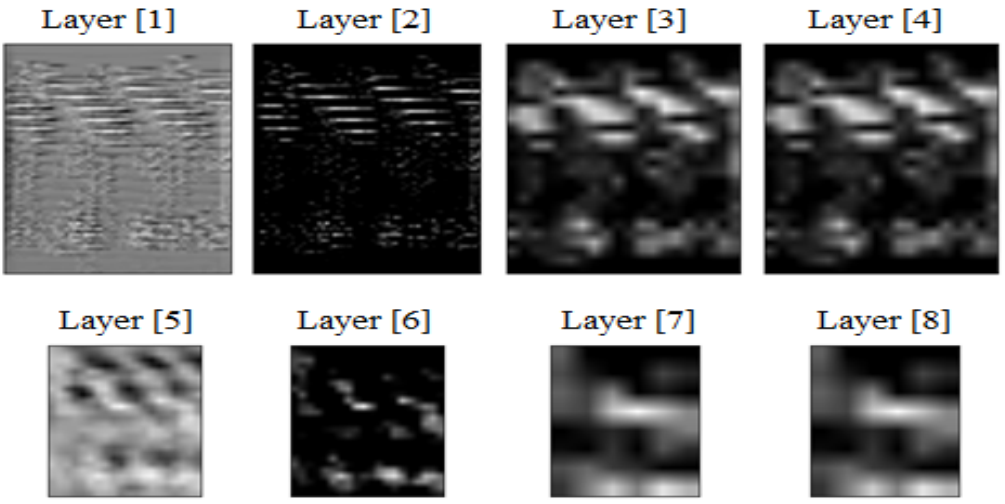


FIGURE 4.4: A feature map image per layer: Convolution layers (Layer 1 and 5), ReLU layers (Layer 2 and 6), Pooling layer (Layer 3 and 7) and Dropout layers (Layer 4,8)

type are reported in Table 4.2. A total number of trainable parameters is 223, 654, which

TABLE 4.2: Model summary for the proposed 2D-CNN model

Layer (type)	Output Shape	Param #
conv2d_1 (Conv2D)	(None, 16, 80, 60)	1024
activation_1 (Activation)	(None, 16, 80, 60)	0
max_pooling2d_1 (MaxPooling2	(None, 16, 26, 20)	0
dropout_1 (Dropout)	(None, 16, 26, 20)	0
conv2d_2 (Conv2D)	(None, 32, 26, 20)	25120
activation_2 (Activation)	(None, 32, 26, 20)	0
max_pooling2d_2 (MaxPooling2	(None, 32, 8, 6)	0
dropout_2 (Dropout)	(None, 32, 8, 6)	0
flatten_1 (Flatten)	(None, 1536)	0
dense_1 (Dense)	(None, 128)	196736
dense_2 (Dense)	(None, 6)	774
activation_3 (Activation)	(None, 6)	0

Total params: 223,654
Trainable params: 223,654
Non-trainable params: 0

is comparatively lower than the one one can get from a model summary of a 3D-CNN for HAR. Since it is trivial that the number of trainable parameters (hyperparameters) has a direct impact on the computational complexity of CNN models, a 2D-CNN presents a lower computational burden as compared to a 3D-CNN model. An exemplary summary for a 3D-CNN model using a simple 3D kernel, justifying the discrepancy in terms of the number of trainable parameters, is depicted in Table 4.3. This summary corresponds to

TABLE 4.3: Model summary for an example of 3D-CNN model

Layer (type)	Output Shape	Param #
conv3d_1 (Conv3D)	(None, 16, 80, 60, 25)	3040
activation_1 (Activation)	(None, 16, 80, 60, 25)	0
max_pooling3d_1 (MaxPooling3	(None, 16, 26, 20, 8)	0
dropout_1 (Dropout)	(None, 16, 26, 20, 8)	0
conv3d_2 (Conv3D)	(None, 32, 26, 20, 8)	75296
activation_2 (Activation)	(None, 32, 26, 20, 8)	0
max_pooling3d_2 (MaxPooling3	(None, 32, 8, 6, 2)	0
dropout_2 (Dropout)	(None, 32, 8, 6, 2)	0
flatten_1 (Flatten)	(None, 3072)	0
dense_1 (Dense)	(None, 128)	393344
dense_2 (Dense)	(None, 6)	774
activation_3 (Activation)	(None, 6)	0

Total params: 472,454
Trainable params: 472,454
Non-trainable params: 0

the extension of the proposed 2D-CNN with a depth equal to 3. The total of trainable parameters amounting to $472,454$ is larger than the one of the proposed 2D-CNN model in Table 4.2. As this amount only considered a video sequence of 25 frames, it would have doubled for a video sequence of 50 frames. Another model summary for a 3D extension of the proposed 2D-CNN model with now a depth of 1 is reported in Table 4.4. The total of trainable parameters for 50 frames considered in a video sequence amounts to $4,942,246$ which is far larger than the amounts of trainable parameters recorded in Tables 4.2 and 4.3. Consequently, the size of dataset contributes in the calculation of the total amount of model parameters which translates into the training computational speed of CNN models.

TABLE 4.4: Model summary for an example of 3D-CNN model

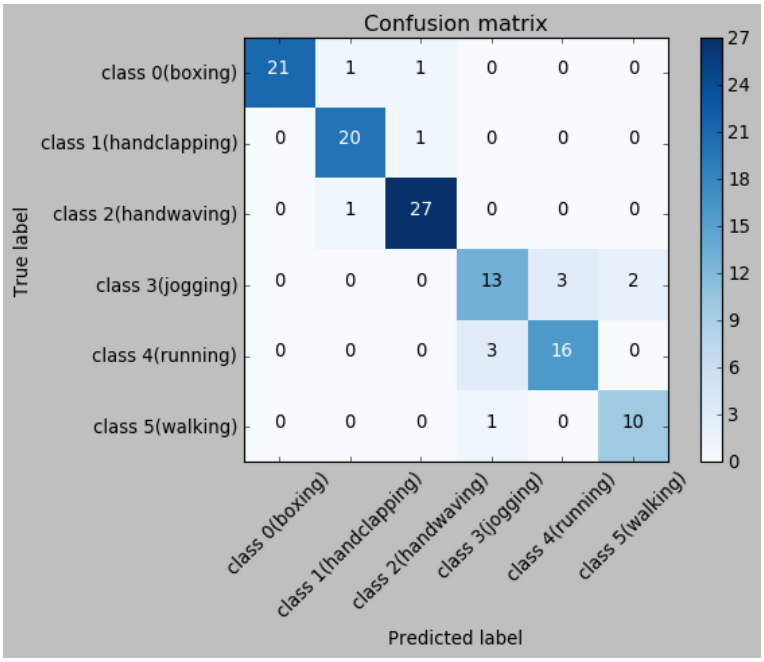
Layer (type)	Output Shape	Param #
conv3d_1 (Conv3D)	(None, 16, 80, 60, 25)	1024
activation_1 (Activation)	(None, 16, 80, 60, 25)	0
max_pooling3d_1 (MaxPooling3	(None, 16, 26, 20, 25)	0
dropout_1 (Dropout)	(None, 16, 26, 20, 25)	0
conv3d_2 (Conv3D)	(None, 32, 26, 20, 25)	25120
activation_2 (Activation)	(None, 32, 26, 20, 25)	0
max_pooling3d_2 (MaxPooling3	(None, 32, 8, 6, 25)	0
dropout_2 (Dropout)	(None, 32, 8, 6, 25)	0
flatten_1 (Flatten)	(None, 38400)	0
dense_1 (Dense)	(None, 128)	4915328
dense_2 (Dense)	(None, 6)	774
activation_3 (Activation)	(None, 6)	0

Total params: 4,942,246
Trainable params: 4,942,246
Non-trainable params: 0

4.3.2 Classification results

The fully connected network represents the classification stage. The output feature maps at layer 8 are both flattened and converted into 128 dimensional feature vectors, then fed into the final layer consisting of six units corresponding to the number of classes (actions). We train our proposed model using the optimization algorithm RMSprop (Root Mean Square prop) that is an adaptive learning rate method. The concept of dropout is also used in a regulatory manner during the learning process to avoid unpleasant overfitting. The classification accuracy across different actions is presented in the form of a confusion matrix shown in Table 4.5. This result corresponds to the testing set representing 20% of the generated MHI database. Misclassification accuracy, though very low, is observed only in actions with the mostly involved body part. Actions such as *Boxing, Handclapping* and *Handwaving* involve mostly hands whereas actions such as *Jogging, Running* and *Walking* involve mostly the feet. This misclassification is realistic and may be due the number of video frames considered and occlusion when computing the MHI templates. One can argue that *jogging*, *running*, and *walking* are similar action but performed at different speeds. In addition, three evaluation metrics, which are precision, recall, and f1-score, are calculated from statistics such as True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) derived from the

TABLE 4.5: Confusion matrix for test recognition accuracy across different actions.



confusion matrix as follows:

$$Precision = \frac{TP}{Tp + FP}, \tag{4.1}$$

$$Recall = \frac{TP}{Tp + FN}, \tag{4.2}$$

and

$$f1 - score = 2 \times \frac{Recall \times Precision}{Recall + Precision}. \tag{4.3}$$

These metrics substantiate the recognition rate obtained for our proposed method and are reported in Table 4.6. Although they are usually used to compare different classifi-

TABLE 4.6: Evaluation metrics for the proposed 2D-CNN model for KTH dataset

action	precision	recall	f1-score	support
class 0 (boxing)	1.00	0.91	0.95	23
class 1 (handclapping)	0.91	0.95	0.93	21
class 2 (handwaving)	0.93	0.96	0.95	28
class 3 (jogging)	0.76	0.72	0.74	18
class 4 (running)	0.84	0.84	0.84	19
class 5 (walking)	0.83	0.91	0.87	11
average/total	0.89	0.89	0.89	120

cation schemes, they can also assist in measuring how good a classification is by means of metric scores for each particular action. The f1-score, also defined as the average of both Precision and Recall metrics, is an important metric to be considered than just

the accuracy metric alone, specially when dealing with unevenly distributed dataset as it is the case here in testing (see the *support* column in Table 4.6). Hence the additional evaluation metrics reported here also reflect the accuracy result, especially with their average values matching the accuracy ones.

Furthermore, the proposed approach achieves an overall accuracy of 89% which is 2% less than the 3D-CNN results obtained in [13]. Our 2D-CNN approach also competes with the 3D-HOG method [2], and it outperforms the handcrafted-based approach developed on the KTH dataset [34]. Table 4.7 shows a comparative performance with the state-of-the-art methods.

TABLE 4.7: Comparative performance accuracies on KTH dataset

Method	Recognition accuracy (%)
3D-CNN [13]	90.20
3D-HOG [2]	91.40
SPS [34]	50.83
Proposed 2D-CNN	**89.17**

4.4 Summary

As previous 2D-CNN approaches were found underperforming for action recognition task due the lack of consideration of the temporal dimension in videos and previous 3D-CNN approaches time found to be time consuming, in this chapter we proposed a 2D-CNN method that included a pre-processing stage generating MHI templates. These templates represent both the spatial and temporal information of videos and were fed to the 2D-CNN to produce salient features overcoming issues of occlusion, self-occlusion, background clutter resulting from MHI images. The proposed 2D-CNN approach was successfully trained and tested. Other evaluation metrics such as precision, recall, and f1-score were computed to prove that our model is satisfactory. The overall results on KTH dataset show competitive performance compared to both the 3D-CNN and 3D-HOG methods. Our results also outperforms the handcrafted based method in [34] by a large margin. It can be concluded that the presented approach achieves recognition rate with a significantly reduced computational complexity imposed by either a 3D-CNN HAR method or a 2D-CNN with the size of dataset.

Chapter 5

Conclusion and future work

"Now this is not the end. It is not even the beginning of the end. But it is, perhaps, the end of the beginning"

Winston Churchill

5.1 Conclusion and contributions

Throughout this thesis, we have been interested in the recognition of human actions forming the basis of many applications such as video surveillance, facial expression recognition, health care, and pedestrian action recognition. To address the problem of recognizing action from visual materials such as images and videos, we explored state-of-the-art holistic approaches and proposed more effective holistic approaches with very low computational complexity.

Chapter 2 provided a comprehensive literature review of previous methods developed by researchers over the past twenty years to address the problem of recognizing human actions. These methods were identified to be either local or holistic, and split by their way of representing actions between handcrafted feature based approaches and deep learning feature based approaches. Although our main focus was on holistic approaches because of their good interpretability, simplicity and computational efficiency, local approaches were briefly introduced for completeness purpose on the topic of human action recognition. From our investigation of several popular holistic approaches alongside with their respective applications and limitations, we discovered that feature extraction and representation as action representation played a crucial role in HAR performance. In addition, handcrafted feature based methods also known as traditional methods relied

on descriptor algorithms that produced suitable feature vectors from reduced database for good performance, whereas the deep learning ones known as modern methods relied on their deep architectures trained with sufficiently large database on powerful hardware to generate robust feature representations for good performance. Both holistic methods were limited by their high computational complexity which remained a great concern. Consequently, our findings suggesting the adoption of existing global feature descriptor or the development of novel ones under a common HAR framework ignited our contributions in the state-of-the-art. These contributions with goal to improve recognition performance in terms of computational complexity were then presented with respect to the approach type in Chapters 3 and 4.

In Chapter 3, proposed handcrafted feature based approaches were summarized in two main contributions. These approaches, apart from treating visual related issues, improved performance by reducing the computational complexity experienced by HAR systems. The first contribution consisted in using the SPS algorithm to extract features of reduced size from MHI templates and fed them into a KNN classifier in order to recognize human actions. Applying the proposed framework on KTH dataset, the per scenario classification results has shown the relevance of the SPS feature as holistic action representation. The second contribution is the design of a novel feature descriptor also referred to as CD–LBP algorithm. This proposed descriptor, inspired by the uniform LBPs which are patterns with at most two circular $0-1$ and $1-0$ transitions, provides higher CD–LBP features as lower dimensional feature vectors which in turn afford a flexibility in making decision on the performance of the recognition system. The relevance of CD–LBP features was then demonstrated in two applications. Each application presented a particular importance in terms of improving the computational complexity of HAR systems. In the first application (FER), higher-order CD–LBP features extracted from the JAFFE dataset in a *histogramming* representation of facial expression resulted in very low dimension feature vectors. This reduction of feature dimensionality highlighted one particularity of the novel descriptor for the higher-order CD–LBP features yielded competitive results as compared to the ones obtained from the original LBP and its variant CS–LBP features in selected state-of-the-art works in terms of both recognition rate and computation speed. The second particularity of the proposed descriptor which is *descriptive* was highlighted in the second application (PAR). The CD–LBP descriptor in conjunction with the HOG descriptor were applied to MHI images resulted from video sequences of pedestrian actions. The highest-order CD–LBP description of MHI image was the most descriptive and suitable for the extraction of high discriminative HOG features. The CD–LBP descriptor impacted more on the improvement of recognition rate than on the computational efficiency. The HOG parameters decided on the dimension of features and control the computational speed

of PAR system. In absence of details on the size of HOG features for action recognition in the literature, we believe that our approach is the simplest and the size provided in this thesis can serve as reference for future research. In addition, we also believe that the binary nature of the highest-order CD–LBP description is advantageous in terms memory usage, especially in a real-time implementation.

In Chapter 4, we proposed a 2D-CNN method that included a pre-processing stage generating MHI templates to recognize human actions. The inclusion of this stage constituted an innovation as it contributed in lowering the computational complexity by reducing the number of hyperparameters. The proposed DL approach was found appropriate for HAR because the number of 2D-CNN operations performed repeatedly on video frames was avoided and the 3D kernels in a 3D-CNN based approach was also prevented. The experimental results on KTH dataset shown comparable performance to those of 3D-CNN method. This approach also outperformed the handchrafted feature based method referred to as SPS method developed in Chapter 3 where SPS descriptor was used as feature extraction technique. Although the proposed 2D-CNN approach produces recognition rate comparable to the state-of-the-art approaches with a significant reduction of the computational complexity in the of the number of hyperparameters, we believe that its recognition rate can be improve by employing MHI templates with less occlusions.

However, our contributions, through all experimental results obtained either for handcrafted or learned features, have proven that the problem of computational complexity of HAR systems can be tackle by simplified action representations.

5.2 Future work

Experimental results of this thesis pave the way for other research perspectives. As the problem of reducing processing times is an ever-pertinent challenge in the recognition of actions from video sequences, it is therefore important to think about possible improvement in computing time without too much compromise on recognition rate. In this section, we then discuss the limitations of proposed holistic approaches and some directions for future work.

First, the recognition performance of the proposed SPS based framework can be improved if issues of occlusion, self-occlusion, and background clutter causing poor MHI templates are addressed. This will probably necessitate the exploration of MHI variants [6] or better background extraction techniques. Moreover, SPS features are not invariant to symmetrical actions as they are captured in only one direction. Therefore, there is

a need to create action representations which are invariant to symmetrical actions such as running to the left and running to the right.

Second, the proposed CD–LBP algorithm can be extended in the spatio-temporal domain. Similarly to LBP-TOP and VLBP descriptors, an 3D extension of CD–LBP descriptor will definitely contribute in the reduction of feature size while achieving a good trade-off between recognition rates and run times. Related MHI issues will then be avoided since in this case the 3D extended CD–LBP descriptor will directly extract features from video sequences.

Third, the performance of the proposed 2D-CNN for HAR, besides the significant reduction of the computational complexity, may be improved if used in combination with the LSTM classifier. Also, based on the fact that the number of hyperparameters has drastically been reduced with the use of MHI templates, we therefore suggest the use of other spatio-temporal templates.

Finally, the proposed holistic approaches can be extended to real-time action recognition. However, this will need observing further real-time requirements which include minimizing the memory demands and computational complexity.

Bibliography

[1] Ronald Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010.

[2] Alexander Klaser, Marcin Marszałek, and Cordelia Schmid. A spatio-temporal descriptor based on 3d-gradients. 2008.

[3] Bappaditya Mandal and How-Lung Eng. Regularized discriminant analysis for holistic human activity recognition. *IEEE Intelligent Systems*, (1):21–31, 2010.

[4] Junji Yamato, Jun Ohya, and Kenichiro Ishii. Recognizing human action in time-sequential images using hidden markov model. In *Proceedings 1992 IEEE Computer Society conference on computer vision and pattern recognition*, pages 379–385. IEEE, 1992.

[5] Alexei A Efros, Alexander C Berg, Greg Mori, and Jitendra Malik. Recognizing action at a distance. In *null*, page 726. IEEE, 2003.

[6] Md Atiqur Rahman Ahad, Joo Kooi Tan, Hyoungseop Kim, and Seiji Ishikawa. Motion history image: its variants and applications. *Machine Vision and Applications*, 23(2):255–281, 2012.

[7] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. *IEEE transactions on pattern analysis and machine intelligence*, 29(12):2247–2253, 2007.

[8] Kellokumpu Vili, Zhao Guoying, and Pietikäinen Matti. Texture based description of movements for activity analysis. In *Int. Conf. on Computer Vision Theory and Applications (VISAPP 2008)*, volume 1, pages 206–213, 2008.

[9] Guoying Zhao and Matti Pietikäinen. Dynamic texture recognition using volume local binary patterns. In *Dynamical Vision*, pages 165–177. Springer, 2006.

[10] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29(6):915–928, 2007.

[11] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[12] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[13] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2012.

[14] Moez Baccouche, Franck Mamalet, Christian Wolf, Christophe Garcia, and Atilla Baskurt. Sequential deep learning for human action recognition. In *International workshop on human behavior understanding*, pages 29–39. Springer, 2011.

[15] Michael Lyons, Shigeru Akamatsu, Miyuki Kamachi, and Jiro Gyoba. Coding facial expressions with gabor wavelets. In *Proceedings Third IEEE international conference on automatic face and gesture recognition*, pages 200–205. IEEE, 1998.

[16] Rifat Muhammad Mueid, Chandrama Ahmed, and Md Atiqur Rahman Ahad. Pedestrian activity classification using patterns of motion and histogram of oriented gradient. *Journal on Multimodal User Interfaces*, 10(4):299–305, 2016.

[17] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36. IEEE, 2004.

[18] Jiang Gao, Alexander G Hauptmann, Ashok Bharucha, and Howard D Wactlar. Dining activity analysis using a hidden markov model. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 2, pages 915–918. IEEE, 2004.

[19] Homa Foroughi, Baharak Shakeri Aski, and Hamidreza Pourreza. Intelligent video surveillance for monitoring fall detection of elderly in home environments. In *2008 11th international conference on computer and information technology*, pages 219–224. IEEE, 2008.

[20] Chin-De Liu, Pau-Choo Chung, Yi-Nung Chung, and Monique Thonnat. Understanding of human behaviors from videos in nursing care monitoring systems. *Journal of High Speed Networks*, 16(1):91–103, 2007.

[21] Ying Luo, Tzong-Der Wu, and Jenq-Neng Hwang. Object-based analysis and interpretation of human motion in sports video sequences by dynamic bayesian networks. *Computer Vision and Image Understanding*, 92(2-3):196–216, 2003.

[22] Khurram Soomro and Amir R Zamir. Action recognition in realistic sports videos. In *Computer vision in sports*, pages 181–208. Springer, 2014.

[23] Manoranjan Paul, Shah ME Haque, and Subrata Chakraborty. Human detection in surveillance videos and its applications-a review. *EURASIP Journal on Advances in Signal Processing*, 2013(1):176, 2013.

[24] Brendan Tran Morris and Mohan Manubhai Trivedi. A survey of vision-based trajectory learning and analysis for surveillance. *IEEE transactions on circuits and systems for video technology*, 18(8):1114–1127, 2008.

[25] Jin Choi, Yong-il Cho, Taewoo Han, and Hyun S Yang. A view-based real-time human action recognition system as an interface for human computer interaction. In *International Conference on Virtual Systems and Multimedia*, pages 112–120. Springer, 2007.

[26] Tongyang Liu, Yang Song, Yu Gu, and Ao Li. Human action recognition based on depth images from microsoft kinect. In *2013 Fourth Global Congress on Intelligent Systems*, pages 200–204. IEEE, 2013.

[27] Maja Pantic, Alex Pentland, Anton Nijholt, and Thomas S Huang. Human computing and machine understanding of human behavior: A survey. In *Artifical Intelligence for Human Computing*, pages 47–71. Springer, 2007.

[28] Matej Hoffmann, Hugo Marques, Alejandro Arieta, Hidenobu Sumioka, Max Lungarella, and Rolf Pfeifer. Body schema in robotics: a review. *IEEE Transactions on Autonomous Mental Development*, 2(4):304–324, 2010.

[29] Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiaris. A review of human activity recognition methods. *Frontiers in Robotics and AI*, 2:28, 2015.

[30] Mariska E Kret. Emotional expressions beyond facial muscle actions. a call for studying autonomic signals and their impact on social perception. *Frontiers in psychology*, 6:711, 2015.

[31] Ivan Laptev. On space-time interest points. *International journal of computer vision*, 64(2-3):107–123, 2005.

[32] Paul Scovanner, Saad Ali, and Mubarak Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 357–360. ACM, 2007.

[33] Fiza Murtaza, Muhammad Haroon Yousaf, and Sergio A Velastin. Multi-view human action recognition using 2d motion templates based on mhis and their hog description. *IET Computer Vision*, 10(7):758–767, 2016.

[34] Ignace Tchangou Toudjeu and Jules Raymond Tapamo. Slope pattern spectra for human action recognition. In *International Conference Image Analysis and Recognition*, pages 381–389. Springer, 2018.

[35] Antonios Oikonomopoulos, Ioannis Patras, and Maja Pantic. Spatiotemporal salient points for visual recognition of human actions. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(3):710–719, 2005.

[36] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[37] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. VS-PETS Beijing, China, 2005.

[38] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *Acm computing surveys (CSUR)*, 38(4):13–es, 2006.

[39] Rajesh Kumar Tripathi, Anand Singh Jalal, and Charul Bhatnagar. A framework for abandoned object detection from video surveillance. In *2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, pages 1–4. IEEE, 2013.

[40] Rajesh Kumar Tripathi, Anand Singh Jalal, and Subhash Chand Agrawal. Suspicious human activity recognition: a review. *Artificial Intelligence Review*, 50(2): 283–339, 2018.

[41] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

[42] Ignace Tchangou Toudjeu and Jules-Raymond Tapamo. Circular derivative local binary pattern feature description for facial expression recognition. *Advances in Electrical and Computer Engineering*, 19(1):51–57, 2019.

[43] Ignace Tchangou Toudjeu, Barend Jacobus van Wyk, Michaël Antonie van Wyk, and Frans van den Bergh. Global image feature extraction using slope pattern spectra. In *International Conference Image Analysis and Recognition*, pages 640–649. Springer, 2008.

[44] Pavan Turaga, Rama Chellappa, Venkatramana S Subrahmanian, and Octavian Udrea. Machine recognition of human activities: A survey. *IEEE Transactions on Circuits and Systems for Video technology*, 18(11):1473, 2008.

[45] Xiaolong Wang, Ali Farhadi, and Abhinav Gupta. Actions˜ transformations. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2658–2667, 2016.

[46] Allah Bux, Plamen Angelov, and Zulfiqar Habib. Vision based human activity recognition: a review. In *Advances in Computational Intelligence Systems*, pages 341–371. Springer, 2017.

[47] Jake K Aggarwal and Quin Cai. Human motion analysis: A review. *Computer vision and image understanding*, 73(3):428–440, 1999.

[48] Xiantong Zhen and Ling Shao. Action recognition via spatio-temporal local features: A comprehensive study. *Image and Vision Computing*, 50:1–13, 2016.

[49] Natalia Díaz Rodríguez, Manuel P Cuéllar, Johan Lilius, and Miguel Delgado Calvo-Flores. A survey on ontologies for human behavior recognition. *ACM Computing Surveys (CSUR)*, 46(4):1–33, 2014.

[50] Jake K Aggarwal and Lu Xia. Human activity recognition from 3d data: A review. *Pattern Recognition Letters*, 48:70–80, 2014.

[51] Fan Zhu, Ling Shao, Jin Xie, and Yi Fang. From handcrafted to learned representations for human action recognition: A survey. *Image and Vision Computing*, 55:42–52, 2016.

[52] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. 2009.

[53] Geert Willems, Tinne Tuytelaars, and Luc Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *European conference on computer vision*, pages 650–663. Springer, 2008.

[54] David G Lowe et al. Object recognition from local scale-invariant features. In *iccv*, volume 99, pages 1150–1157, 1999.

[55] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

[56] Xiantong Zhen and Ling Shao. Introduction to human action recognition. *Wiley Encyclopedia of Electrical and Electronics Engineering*, pages 1–11, 1999.

[57] Dao-Qing Dai and Pong C Yuen. Face recognition by regularized discriminant analysis. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 37(4):1080–1085, 2007.

[58] Gül Ahmet Bahtiyar et al. Holistic face recognition by dimension reduction. 2003.

[59] Xudong Jiang, Bappaditya Mandal, and Alex Kot. Eigenfeature regularization and extraction in face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):383–394, 2008.

[60] Muhammad Murtaza Khan, Muhammad Younus Javed, and Muhammad Almas Anjum. Face recognition using sub-holistic pca. In *2005 International Conference on Information and Communication Technologies*, pages 152–157. IEEE, 2005.

[61] Vandana S Bhat and Jagadeesh D Pujari. Face recognition using holistic based approach. *International Journal of Emerging Technology and Advanced Engineering*, 4(7):134–141, 2014.

[62] Wenyi Zhao, Arvindh Krishnaswamy, Rama Chellappa, Daniel L Swets, and John Weng. Discriminant analysis of principal components for face recognition. In *Face Recognition*, pages 73–85. Springer, 1998.

[63] Kyungnam Kim. Face recognition using principle component analysis. In *International Conference on Computer Vision and Pattern Recognition*, volume 586, page 591, 1996.

[64] Liton Chandra Paul and Abdulla Al Sumam. Face recognition using principal component analysis method. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 1(9):135–139, 2012.

[65] Peter N Belhumeur, João P Hespanha, and David J Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):711–720, 1997.

[66] E Jyotsna, PV Akhil, and Arun Kumar. Silhouette based human action recognition using pca and isomap. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(11), 2013.

[67] A Jeyanthi Suresh and P Asha. Human action recognition in video using histogram of oriented gradient (hog) features and probabilistic neural network (pnn). *International Journal of Innovative Research in Computer and Communication Engineering*, 4(7), 2016.

[68] Riccardo Mattivi and Ling Shao. Human action recognition using lbp-top as sparse spatio-temporal feature descriptor. In *International Conference on Computer Analysis of Images and Patterns*, pages 740–747. Springer, 2009.

[69] Honghai Liu, Zhaojie Ju, Xiaofei Ji, Chee Seng Chan, and Mehdi Khoury. Study of human action recognition based on improved spatio-temporal features. In *Human Motion Sensing and Recognition*, pages 233–250. Springer, 2017.

[70] Alok Sharma and Kuldip K Paliwal. Linear discriminant analysis for the small sample size problem: an overview. *International Journal of Machine Learning and Cybernetics*, 6(3):443–454, 2015.

[71] Md Zia Uddin, JJ Lee, and T-S Kim. Independent component feature-based human activity recognition via linear discriminant analysis and hidden markov model. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5168–5171. IEEE, 2008.

[72] Angkoon Phinyomark, Huosheng Hu, Pornchai Phukpattaranont, and Chusak Limsakul. Application of linear discriminant analysis in dimensionality reduction for hand motion classification. *Measurement Science Review*, 12(3):82–89, 2012.

[73] Ming Guo and Zhelong Wang. A feature extraction method for human action recognition using body-worn inertial sensors. In *2015 IEEE 19th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 576–581. IEEE, 2015.

[74] Itzik Pima and Mayer Aladjem. Regularized discriminant analysis for face recognition. *Pattern Recognition*, 37(9):1945–1948, 2004.

[75] Jianhua Zhao, Lei Shi, and Ji Zhu. Two-stage regularized linear discriminant analysis for 2-d data. *IEEE transactions on neural networks and learning systems*, 26(8):1669–1681, 2014.

[76] Juwei Lu, Konstantinos N Plataniotis, and Anastasios N Venetsanopoulos. Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition. *Pattern recognition letters*, 26(2):181–191, 2005.

[77] Juwei Lu, Konstantinos N Plataniotis, and Anastasios N Venetsanopoulos. Face recognition using lda-based algorithms. *IEEE Transactions on Neural networks*, 14(1):195–200, 2003.

[78] Jerome H Friedman. Regularized discriminant analysis. *Journal of the American statistical association*, 84(405):165–175, 1989.

[79] Bappaditya Mandal and How-Lung Eng. 3-parameter based eigenfeature regularization for human activity recognition. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 954–957. IEEE, 2010.

[80] Festus Osayamwen and Jules R Tapamo. Within-class subspace regularization for human activity recognition. In *2016 IEEE International Conference on Industrial Technology (ICIT)*, pages 804–807. IEEE, 2016.

[81] Aaron F Bobick and James W Davis. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (3):257–267, 2001.

[82] David Nicholas Olivieri, Iván Gómez Conde, and Xosé Antón Vila Sobrino. Eigenspace-based fall detection and activity recognition from motion templates and machine learning. *Expert Systems with Applications*, 39(5):5935–5945, 2012.

[83] Qiang Zhu, Mei-Chen Yeh, Kwang-Ting Cheng, and Shai Avidan. Fast human detection using a cascade of histograms of oriented gradients. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1491–1498. IEEE, 2006.

[84] WSK Fernando, HMSPB Herath, PH Perera, MPB Ekanayake, GMRI Godaliyadda, and JV Wijayakulasooriya. Object identification, enhancement and tracking under dynamic background conditions. In *7th International Conference on Information and Automation for Sustainability*, pages 1–6. IEEE, 2014.

[85] S Sulaiman, Aini Hussain, N Tahir, Salina Abdul Samad, and Mohd Marzuki Mustafa. Human silhouette extraction using background modeling and subtraction techniques. *Information Technology Journal*, 7(1):155–159, 2008.

[86] Jeisung Lee and Mignon Park. An adaptive background subtraction method based on kernel density estimation. *Sensors*, 12(9):12279–12300, 2012.

[87] B Lee and M Hedley. Background estimation for video surveillance. 2002.

[88] Massimo Piccardi. Background subtraction techniques: a review. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, volume 4, pages 3099–3104. IEEE, 2004.

[89] Stefano Messelodi, Carla Maria Modena, Nicola Segata, and Michele Zanin. A kalman filter based background updating algorithm robust to sharp illumination changes. In *International Conference on Image Analysis and Processing*, pages 163–170. Springer, 2005.

[90] Mohamad Hoseyn Sigari, Naser Mozayani, and H Pourreza. Fuzzy running average and fuzzy background subtraction: concepts and application. *International Journal of Computer Science and Network Security*, 8(2):138–143, 2008.

[91] Shah Atiqur Rahman, S-Y Cho, and MKH Leung. Recognising human actions by analysing negative spaces. *IET Computer Vision*, 6(3):197–213, 2012.

[92] Daniel Weinland, Edmond Boyer, and Remi Ronfard. Action recognition from arbitrary views using 3d exemplars. 2007.

[93] Dinesh Kumar Vishwakarma and Rajiv Kapoor. Hybrid classifier based human activity recognition using the silhouette and cells. *Expert Systems with Applications*, 42(20):6957–6965, 2015.

[94] Alexandros Andre Chaaraoui and Francisco Flórez-Revuelta. A low-dimensional radial silhouette-based feature for fast human action recognition fusing multiple views. *International scholarly research notices*, 2014, 2014.

[95] Alexandros Andre Chaaraoui, Pau Climent-Pérez, and Francisco Flórez-Revuelta. Silhouette-based human action recognition using sequences of key poses. *Pattern Recognition Letters*, 34(15):1799–1807, 2013.

[96] Ramprasad Polana and Randal Nelson. Low level recognition of human motion (or how to get your man without finding his body parts). In *Proceedings of 1994 IEEE Workshop on Motion of Non-rigid and Articulated Objects*, pages 77–82. IEEE, 1994.

[97] Alireza Fathi and Greg Mori. Action recognition by learning mid-level motion features. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[98] Yang Wang, Payam Sabzmeydani, and Greg Mori. Semi-latent dirichlet allocation: A hierarchical model for human action recognition. In *Workshop on Human Motion*, pages 240–254. Springer, 2007.

[99] Neil Robertson and Ian Reid. Behaviour understanding in video: a combined method. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 808–815. IEEE, 2005.

[100] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981.

[101] Hanno Scharr. Optimal filters for extended optical flow. In *International Workshop on Complex Motion*, pages 14–29. Springer, 2004.

[102] Ashok Ramadass, Myunghoon Suk, and Balakrishnan Prabhakaran. Feature extraction method for video based human action recognitions: extended optical flow algorithm. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1106–1109. IEEE, 2010.

[103] Aaron F Bobick. Movement, activity and action: the role of knowledge in the perception of motion. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 352(1358):1257–1265, 1997.

[104] Thomas B Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer vision and image understanding*, 104(2-3):90–126, 2006.

[105] Aaron F Bobick and James W Davis. Action recognition using temporal templates. In *Motion-Based Recognition*, pages 125–146. Springer, 1997.

[106] Md Atiqur Rahman Ahad. *Motion history images for action recognition and understanding*. Springer Science & Business Media, 2012.

[107] Md Atiqur Rahman Ahad, Md Nazmul Islam, and Israt Jahan. Action recognition based on binary patterns of action-history and histogram of oriented gradient. *Journal on Multimodal User Interfaces*, 10(4):335–344, 2016.

[108] Zhaozheng Yin and Robert Collins. Moving object localization in thermal imagery by forward-backward mhi. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, pages 133–133. IEEE, 2006.

[109] Moshe Blank, Lena Gorelick, Eli Shechtman, Michal Irani, and Ronen Basri. Actions as space-time shapes. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1395–1402. IEEE, 2005.

[110] Daniel Weinland, Remi Ronfard, and Edmond Boyer. Free viewpoint action recognition using motion history volumes. *Computer vision and image understanding*, 104(2-3):249–257, 2006.

[111] Daniel Weinland, Remi Ronfard, and Edmond Boyer. Automatic discovery of action taxonomies from multiple views. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1639–1645. IEEE, 2006.

[112] Hossein Ragheb, Sergio Velastin, Paolo Remagnino, and Tim Ellis. Human action recognition using robust power spectrum features. In *2008 15th IEEE International Conference on Image Processing*, pages 753–756. IEEE, 2008.

[113] Alexandra Branzan Albu and Trevor Beugeling. A three-dimensional spatiotemporal template for interactive human motion analysis. *Journal of Multimedia*, 2 (4), 2007.

[114] Mohiuddin Ahmad, Irine Parvin, and Seong-Whan Lee. Silhouette history and energy image information for human movement recognition. *Journal of Multimedia*, 5(1), 2010.

[115] Varsha H Chandrashekhar and KS Venkatesh. Action energy images for reliable human action recognition. In *Proceedings of the Asian Symposium on Information Display (ASID 2006)*, pages 484–487. Citeseer, 2006.

[116] Liang Wang and David Suter. Informative shape representations for human action recognition. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 2, pages 1266–1269. IEEE, 2006.

[117] Romer Rosales. Recognition of human action using moment-based feature. Technical report, Citeseer, 1998.

[118] Md Atiqur Rahman Ahad, Joo Kooi Tan, HS Kim, and Seiji Ishikawa. Temporal motion recognition and segmentation approach. *International Journal of Imaging Systems and Technology*, 19(2):91–99, 2009.

[119] Md Atiqur Rahman Ahad, Joo Kooi Tan, Hyoungseop Kim, and Seiji Ishikawa. Analysis of motion self-occlusion problem due to motion overwriting for human activity recognition. *Journal of Multimedia*, 5(1):36–46, 2010.

[120] Hongying Meng, Nick Pears, and Chris Bailey. A human action recognition system for embedded computer vision application. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6. IEEE, 2007.

[121] R Venkatesh Babu and Kalpathi R Ramakrishnan. Recognition of human actions using motion history information extracted from the compressed video. *Image and Vision computing*, 22(8):597–607, 2004.

[122] Shiv N Vitaladevuni, Vili Kellokumpu, and Larry S Davis. Action recognition using ballistic dynamics. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008.

[123] Caifeng Shan, Yucheng Wei, Xianchao Qiu, and Tieniu Tan. Gesture recognition using temporal template based trajectories. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 954–957. IEEE, 2004.

[124] Karianto Leman, Goel Ankit, and Tele Tan. Pda based human motion recognition system. *International Journal of Software Engineering and Knowledge Engineering*, 15(02):199–204, 2005.

[125] W Ryu, D Kim, HS Lee, J Sung, and D Kim. Gesture recognition using temporal templates. *Proc. ICPR, Demo Program, Hong Kong*, 2006.

[126] Ju Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE transactions on pattern analysis and machine intelligence*, 28(2):316–322, 2005.

[127] Xiaotao Zou and Bir Bhanu. Human activity classification based on gait energy image and coevolutionary genetic programming. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 556–559. IEEE, 2006.

[128] Changhong Chen, Jimin Liang, Heng Zhao, Haihong Hu, and Jie Tian. Frame difference energy image for gait recognition with incomplete silhouettes. *Pattern Recognition Letters*, 30(11):977–984, 2009.

[129] Truong Thien Dinh, Tran Thi Bich Hanh, Ho Ngoc Lam, et al. Detection and localization of road area in traffic video sequences using motion information and fuzzy-shadowed sets. In *Seventh IEEE International Symposium on Multimedia (ISM'05)*, pages 6–pp. IEEE, 2005.

[130] James W Davis, Alexander M Morison, and David D Woods. Building adaptive camera models for video surveillance. In *2007 IEEE Workshop on Applications of Computer Vision (WACV'07)*, pages 34–34. IEEE, 2007.

[131] Alexia Briassouli and Ioannis Kompatsiaris. Robust temporal activity templates using higher order statistics. *IEEE Transactions on Image Processing*, 18(12): 2756–2768, 2009.

[132] Hongying Meng, Nick Pears, Michael Freeman, and Chris Bailey. Motion history histograms for human action recognition. In *Embedded Computer Vision*, pages 139–162. Springer, 2009.

[133] Chin-Pan Huang, Chaur-Heh Hsieh, Kuan-Ting Lai, and Wei-Yang Huang. Human action recognition using histogram of oriented gradient of motion history image. In *2011 First International Conference on Instrumentation, Measurement, Computer, Communication and Control*, pages 353–356. IEEE, 2011.

[134] K Akila and S Chitrakala. An efficient method to resolve intraclass variability using highly refined hog description model for human action recognition. *CONCURRENCY AND COMPUTATION-PRACTICE & EXPERIENCE*, 31(12), 2019.

[135] Xiaodong Yang, Chenyang Zhang, and YingLi Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1057–1060, 2012.

[136] Yuanyuan Huang, Haomiao Yang, and Ping Huang. Action recognition using hog feature in different resolution video sequences. In *2012 International Conference on Computer Distributed Control and Intelligent Environmental Monitoring*, pages 85–88. IEEE, 2012.

[137] Vili Kellokumpu, Guoying Zhao, and Matti Pietikäinen. Human activity recognition using a dynamic texture based method. In *BMVC*, volume 1, page 2, 2008.

[138] Vili Kellokumpu, Guoying Zhao, and Matti Pietikäinen. Recognition of human actions using texture descriptors. *Machine Vision and Applications*, 22(5):767–780, 2011.

[139] Lahav Yeffet and Lior Wolf. Local trinary patterns for human action recognition. In *2009 IEEE 12th international conference on computer vision*, pages 492–497. IEEE, 2009.

[140] Carlos Orrite, Mario Rodriguez, Elias Herrero, Gregory Rogez, and Sergio A Velastin. Automatic segmentation and recognition of human actions in monocular sequences. In *2014 22nd International Conference on Pattern Recognition*, pages 4218–4223. IEEE, 2014.

[141] Sanchit Singh, Sergio A Velastin, and Hossein Ragheb. Muhavi: A multicamera human action video dataset for the evaluation of action recognition methods. In *2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 48–55. IEEE, 2010.

[142] Shahzad Cheema, Abdalrahman Eweiwi, Christian Thurau, and Christian Bauckhage. Action recognition by learning discriminative key poses. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1302–1309. IEEE, 2011.

[143] PA Dhulekar and ST Gandhe. Action recognition based on histogram of oriented gradients and spatio-temporal interest points. *International Journal of Engineering & Technology*, 7(4):2153–2160, 2018.

[144] Nguyen Thanh Binh, Swati Nigam, and Ashish Khare. Towards classification based human activity recognition in video sequences. In *International Conference on Context-Aware Systems and Applications*, pages 209–218. Springer, 2013.

[145] Matti Pietikäinen, Abdenour Hadid, Guoying Zhao, and Timo Ahonen. *Computer vision using local binary patterns*, volume 40. Springer Science & Business Media, 2011.

[146] Caroline Silva, Thierry Bouwmans, and Carl Frélicot. An extended center-symmetric local binary pattern for background modeling and subtraction in videos. 2015.

[147] Di Huang, Caifeng Shan, Mohsen Ardabilian, Yunhong Wang, and Liming Chen. Local binary patterns and its application to facial image analysis: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 41(6):765–781, 2011.

[148] Xiaoyi Feng, M Pietikainen, and Abdenour Hadid. Facial expression recognition with local binary patterns and linear programming. *Pattern Recognition And Image Analysis C/C of Raspoznavaniye Obrazov I Analiz Izobrazhenii*, 15(2):546, 2005.

[149] SL Happy, Anjith George, and Aurobinda Routray. A real time facial expression classification system using local binary patterns. In *2012 4th International conference on intelligent human computer interaction (IHCI)*, pages 1–5. IEEE, 2012.

[150] Xiaoming Zhao and Shiqing Zhang. A review on facial expression recognition: feature extraction and classification. *IETE Technical Review*, 33(5):505–517, 2016.

[151] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[152] Topi Mäenpää and Matti Pietikäinen. Texture analysis with local binary patterns. In *Handbook of pattern recognition and computer vision*, pages 197–216. World Scientific, 2005.

[153] Vili Kellokumpu, Guoying Zhao, and Matti Pietikäinen. Dynamic textures for human movement recognition. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pages 470–476, 2010.

[154] Han Su, Jiayun Zou, and Wenjie Wang. Human activity recognition based on silhouette analysis using local binary patterns. In *2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, pages 924–929. IEEE, 2013.

[155] Li Deng, Dong Yu, et al. Deep learning: methods and applications. *Foundations and Trends® in Signal Processing*, 7(3–4):197–387, 2014.

[156] Yann LeCun, Fu Jie Huang, and Leon Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *Proceedings of the 2004 IEEE*

Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004., volume 2, pages II–104. IEEE, 2004.

[157] Wanli Ouyang, Xingyu Zeng, Xiaogang Wang, Shi Qiu, Ping Luo, Yonglong Tian, Hongsheng Li, Shuo Yang, Zhe Wang, Hongyang Li, et al. Deepid-net: Object detection with deformable part based convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(7):1320–1334, 2016.

[158] Tarik A Rashid. Convolutional neural networks based method for improving facial expression recognition. In *The International Symposium on Intelligent Systems Technologies and Applications*, pages 73–84. Springer, 2016.

[159] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014.

[160] Liang Lin, Keze Wang, Wangmeng Zuo, Meng Wang, Jiebo Luo, and Lei Zhang. A deep structured model with radius–margin bound for 3d human activity recognition. *International Journal of Computer Vision*, 118(2):256–273, 2016.

[161] Chi Geng and JianXin Song. Human action recognition based on convolutional neural networks with a convolutional auto-encoder. In *2015 5th International Conference on Computer Sciences and Automation Engineering (ICCSAE 2015)*. Atlantis Press, 2016.

[162] J Arunnehru, G Chamundeeswari, and S Prasanna Bharathi. Human action recognition using 3d convolutional neural networks with 3d motion cuboids in surveillance videos. *Procedia computer science*, 133:471–477, 2018.

[163] Lei Wang, Yangyang Xu, Jun Cheng, Haiying Xia, Jianqin Yin, and Jiaji Wu. Human action recognition by learning spatio-temporal features with deep neural networks. *IEEE access*, 6:17913–17922, 2018.

[164] Saleh Albelwi and Ausif Mahmood. A framework for designing the architectures of deep convolutional neural networks. *Entropy*, 19(6):242, 2017.

[165] Yuzhen Lu and Renfu Lu. Detection of surface and subsurface defects of apples using structured-illumination reflectance imaging with machine learning algorithms. *Transactions of the ASABE*, 61(6):1831–1842, 2018.

[166] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[167] Dominik Scherer, Andreas Müller, and Sven Behnke. Evaluation of pooling operations in convolutional architectures for object recognition. In *International conference on artificial neural networks*, pages 92–101. Springer, 2010.

[168] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[169] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[170] Léon Bottou. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes*, 91(8):12, 1991.

[171] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

[172] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[173] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(Jul):2121–2159, 2011.

[174] Léon Bottou. Stochastic gradient descent tricks. In *Neural networks: Tricks of the trade*, pages 421–436. Springer, 2012.

[175] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[176] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[177] Xiaojiang Peng, Limin Wang, Xingxing Wang, and Yu Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *Computer Vision and Image Understanding*, 150:109–125, 2016.

[178] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.

[179] Lin Sun, Kui Jia, Dit-Yan Yeung, and Bertram E Shi. Human action recognition using factorized spatio-temporal convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4597–4605, 2015.

[180] Keze Wang, Xiaolong Wang, Liang Lin, Meng Wang, and Wangmeng Zuo. 3d human activity recognition with reconfigurable convolutional neural networks. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 97–106, 2014.

[181] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[182] Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1510–1517, 2017.

[183] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.

[184] Hao Yang, Chunfeng Yuan, Bing Li, Yang Du, Junliang Xing, Weiming Hu, and Stephen J Maybank. Asymmetric 3d convolutional neural networks for action recognition. *Pattern recognition*, 85:1–12, 2019.

[185] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[186] Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks*, 5(2):157–166, 1994.

[187] Abdelhadi Azzouni and Guy Pujolle. A long short-term memory recurrent neural network framework for network traffic matrix prediction. *arXiv preprint arXiv:1705.05690*, 2017.

[188] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[189] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4694–4702, 2015.

[190] Qing Li, Zhaofan Qiu, Ting Yao, Tao Mei, Yong Rui, and Jiebo Luo. Action recognition by learning deep multi-granular spatio-temporal video representation. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, pages 159–166, 2016.

[191] Huafeng Chen, Jun Chen, Ruimin Hu, Chen Chen, and Zhongyuan Wang. Action recognition with temporal scale-invariant deep learning framework. *China Communications*, 14(2):163–172, 2017.

[192] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[193] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European conference on computer vision*, pages 816–833. Springer, 2016.

[194] Juha Karhunen, Tapani Raiko, and KyungHyun Cho. Unsupervised deep learning: A short review. In *Advances in Independent Component Analysis and Learning Machines*, pages 125–142. Elsevier, 2015.

[195] Miguel A Carreira-Perpinan and Geoffrey E Hinton. On contrastive divergence learning. In *Aistats*, volume 10, pages 33–40. Citeseer, 2005.

[196] Bo Chen, Jo-Anne Ting, Ben Marlin, and Nando de Freitas. Deep learning of invariant spatio-temporal features from video. In *NIPS 2010 Deep Learning and Unsupervised Feature Learning Workshop*, 2010. URL http://www.cs.ubc.ca/~nando/papers/nipsworkshop2010.pdf.

[197] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554, 2006.

[198] Pasquale Foggia, Alessia Saggese, Nicola Strisciuglio, and Mario Vento. Exploiting the deep learning paradigm for recognizing human actions. In *2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 93–98. IEEE, 2014.

[199] Hülya Yalçın. Human activity recognition using deep belief networks. In *2016 24th Signal processing and communication application conference (SIU)*, pages 1649–1652. IEEE, 2016.

[200] Ruslan Salakhutdinov and Geoffrey Hinton. Deep boltzmann machines. In *Artificial intelligence and statistics*, pages 448–455, 2009.

[201] Ruslan Salakhutdinov and Hugo Larochelle. Efficient learning of deep boltzmann machines. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 693–700, 2010.

[202] Bo Chen. *Deep learning of invariant spatio-temporal features from video.* PhD thesis, University of British Columbia, 2010.

[203] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning*, pages 609–616, 2009.

[204] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Unsupervised learning of hierarchical representations with convolutional deep belief networks. *Communications of the ACM*, 54(10):95–103, 2011.

[205] Gary B Huang, Honglak Lee, and Erik Learned-Miller. Learning hierarchical representations for face verification with convolutional deep belief networks. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2518–2525. IEEE, 2012.

[206] Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

[207] Arun Sharma, S Kumar, N McLachalan, et al. Representation and classification of human movement using temporal templates and statistical measure of similarity. In *WITSP'2002 Workshop on Internet Telecommunications and Signal Processing*, pages 191–196, 2002.

[208] Adem Karahoca and Murat Nurullahoglu. Human motion analysis and action recognition. In *WSEAS International Conference. Proceedings. Mathematics and Computers in Science and Engineering*, volume 7. World Scientific and Engineering Academy and Society, 2008.

[209] N Vapnik Vladimir and V Vapnik. Statistical learning theory. *Xu JH and Zhang XG. translation. Beijing: Publishing House of Electronics Industry, 2O04*, 1998.

[210] Md Ismail Hossain Raju, Sharmeen Sultana Ananna, Syed Shafiul Islam Meraz, Md Zakaria Azam, Seiichi Serikawa, and Md Atiqur Rahman Ahad. Human action recognition: a template matching-based approach. *J Inst Ind Appl Eng*, 5(1):15–23, 2017.

[211] Ferdous Ali Israt, Mostafa Zaman, Mosabber Uddin Ahmed, Syoji Kobashi, and Md Atiqur Rahman Ahad. A study on human action recognition based on a modified-mhi. *International Journal of Biomedical Soft Computing and Human Sciences: the official journal of the Biomedical Fuzzy Systems Association*, 23(1): 37–50, 2018.

[212] Tao Xiang and Shaogang Gong. Beyond tracking: Modelling activity and understanding behaviour. *International Journal of Computer Vision*, 67(1):21–51, 2006.

[213] Hakan Bilen, Basura Fernando, Efstratios Gavves, and Andrea Vedaldi. Action recognition with dynamic image networks. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2799–2813, 2017.

[214] Nidhi N Khatri, Zankhana H Shah, and Samip A Patel. Facial expression recognition: A survey. *International Journal of Computer Science and Information Technologies (IJCSIT)*, 5(1):149–152, 2014.

[215] Anima Majumder, Laxmidhar Behera, and Venkatesh K Subramanian. Local binary pattern based facial expression recognition using self-organizing map. In *2014 International Joint Conference on Neural Networks (IJCNN)*, pages 2375–2382. IEEE, 2014.

[216] Gengjian Xue, Li Song, Jun Sun, and Meng Wu. Hybrid center-symmetric local pattern for dynamic background subtraction. In *2011 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2011.

[217] Olli Lahdenoja, Jonne Poikonen, and Mika Laiho. Towards understanding the formation of uniform local binary patterns. *ISRN Machine Vision*, 2013, 2013.

[218] Caifeng Shan, Shaogang Gong, and Peter W McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and vision Computing*, 27(6):803–816, 2009.

[219] Stephen Moore and Richard Bowden. Local binary patterns for multi-view facial expression recognition. *Computer vision and image understanding*, 115(4):541–558, 2011.

[220] Ira Cohen, Nicu Sebe, Ashutosh Garg, Michael S Lew, and Thomas S Huang. Facial expression recognition from video sequences. In *Proceedings. IEEE International Conference on Multimedia and Expo*, volume 2, pages 121–124. IEEE, 2002.

[221] Abu Sayeed Md Sohail and Prabir Bhattacharya. Classification of facial expressions using k-nearest neighbor classifier. In *International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications*, pages 555–566. Springer, 2007.

[222] R Suresh, S Audithan, G Kannan, and K Raja. Facial expression recognition system using local texture features of contourlet transformation. *Australian Journal of Basic and Applied Sciences*, 10(2), 2016.

[223] Sea Kasim, Rohayanti Hassan, Nur Hadiana Zaini, AS Ahmad, Azizul Azhar Ramli, and Rd Rohmat Saedudin. A study on facial expression recognition using local binary pattern. *Int. J. Adv. Sci. Eng. Inf. Technol.*, 7(5):1621–1626, 2017.

[224] Ya Chang, Changbo Hu, Rogerio Feris, and Matthew Turk. Manifold based analysis of facial expression. *Image and Vision Computing*, 24(6):605–614, 2006.

[225] Stefano Berretti, Alberto Del Bimbo, Pietro Pala, Boulbaba Ben Amor, and Mohamed Daoudi. A set of selected sift features for 3d facial expression recognition. In *2010 20th International Conference on Pattern Recognition*, pages 4125–4128. IEEE, 2010.

[226] Xiaoming Zhao and Shiqing Zhang. Facial expression recognition using local binary patterns and discriminant kernel locally linear embedding. *EURASIP journal on Advances in signal processing*, 2012(1):20, 2012.

[227] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[228] Michael J Lyons, Julien Budynek, and Shigeru Akamatsu. Automatic classification of single facial images. *IEEE transactions on pattern analysis and machine intelligence*, 21(12):1357–1362, 1999.

[229] Xiaoyi Feng, Baohua Lv, Zhen Li, and Jiling Zhang. A novel feature extraction method for facial expression recognition. In *9th Joint International Conference on Information Sciences (JCIS-06)*. Atlantis Press, 2006.

[230] K Meena and A Suruliandi. Local binary patterns and its variants for face recognition. In *2011 International Conference on Recent Trends in Information Technology (ICRTIT)*, pages 782–786. IEEE, 2011.

[231] Yao Wu and Weigen Qiu. Facial expression recognition based on improved deep belief networks. In *AIP Conference Proceedings*, volume 1864, page 020130. AIP Publishing LLC, 2017.

[232] Young-Nam Kim, Jin-Hee Park, and Moon-Hyun Kim. Spatio-temporal analysis of trajectory for pedestrian activity recognition. *Journal of Electrical Engineering & Technology*, 13(2):961–968, 2018.

[233] Li Chen, Nan Ma, Patrick Wang, Jiahong Li, Pengfei Wang, Guilin Pang, and Xiaojun Shi. Survey of pedestrian action recognition techniques for autonomous driving. *Tsinghua Science and Technology*, 25(4):458–470, 2020.

[234] Xiaoyu Wang, Tony X Han, and Shuicheng Yan. An HOG–LBP human detector with partial occlusion handling. In *2009 IEEE 12th international conference on computer vision*, pages 32–39. IEEE, 2009.