The University Of Sheffield.

Privacy aware human action recognition: An exploration of temporal salience modelling and neuromorphic vision sensing

Salah Mahdi Saleh Al-Obaidi

Supervisor:

Dr. Charith Abhayaratne

Thesis submitted in candidature for graduating with a degree of doctor of philosophy from the University of Sheffield, Department of Electronic and Electrical Engineering, Faculty of Engineering, March 14, 2020

Abstract

Solving the issue of privacy in the application of vision-based home monitoring has emerged as a significant demand. The state-of-the-art studies contain advanced privacy protection by filtering/covering the most sensitive content, which is the identity in this scenario. However, going beyond privacy remains a challenge for the machine to explore the obfuscated data, *i.e.*, utility. Thanks for the usefulness of exploring the human visual system to solve the problem of visual data. Nowadays, a high level of visual abstraction can be obtained from the visual scene by constructing saliency maps that highlight the most useful content in the scene and attenuate others. One way of maintaining privacy with keeping useful information about the action is by discovering the most significant region and removing the redundancy. Another solution to address the privacy is motivated by the new visual sensor technology, *i.e.*, neuromorphic vision sensor. In this thesis, we first introduce a novel method for vision-based privacy preservation. Particularly, we propose a new temporal salience-based anonymisation method to preserve privacy with maintaining the usefulness of the anonymity domain-based data. This anonymisation method has achieved a high level of privacy compared to the current work. The second contribution involves the development of a new descriptor for human action recognition (HAR) based on exploring the anonymity domain of the temporal salience method. The proposed descriptor tests the utility of the anonymised data without referring to RGB intensities of the original data. The extracted features using our proposed descriptor have shown an improvement with accuracies of the human actions, outperforming the existing methods. The proposed method has shown improvements by 3.04%, 3.14%, 0.83%, 3.67%, and 16.71% for DHA, KTH, UIUC1, UCF sports, and HMDB51 datasets, respectively, compared to state-of-the-art methods. The third contribution focuses on proposing a new method to deal with the new neuromorphic vision domain, which has come up to the application, since the issue of privacy has been already solved by the sensor itself. The output of this new domain is exploited by further exploring the local and global details of the log intensity changes. The empirical evaluation shows that exploring the neuromorphic domain provides useful details that have demonstrated increasing accuracy rates for E-KTH, E-UCF11 and E-HMDB5 by 0.54%, 19.42% and 25.61%, respectively.

Contents

List of Figures

List of Tables

List of symbols

β	The block dimension of 2DFFT
τ	The intensity deference threshold
$\tilde{\mathcal{E}}$	The spectral entropy
$\hat{\mathcal{E}}$	The weighted entropy
$\mathcal{N}$	The number of thresholds
$\mathcal{A}_{b_m}$	The magnitude of the block
θ_t	The orientation of the gradient of the temporal salience map
$\mathcal{L}$	The log intensity
Θ	The temporal contrast threshold
$\mathbb{E}$	A stream of events
$\mathbb{T}$	The set of all time intervals of a stream of events
$\mathcal{T}_w$	The set of events in a time interval
$\mathcal{G}$	The number of groups including the successive events
ρ	The number of the successive events
μ_x	Mean of x coordinate
μ_y	Mean of y coordinate
γ	The skewness
κ	kurtosis
σ^2	Standard deviation
σ	Variance
$\mathcal{F}_S$	The local feature vector
$\mathcal{F}_T$	The global feature vector
$\mathbb{H}$	The histogram of run-length of polarities
$\mathcal{H}_{\mathbb{E}}$	The temporal groups of events at a specific location
δ_h	Temporal group of events
$\mathbb{F}(\mathbb{E})$	The local and global temporal feature vector

A_d	The total correct answers
Av	The average of the correct answers
b_m	The block m in the saliency region
B	The set of blocks in the salient region
C	Action dataset
$C_{\ell_{(x,y)}}$	The number of events recorded on each of nine spatial coordinates
D	The difference map before thresholding
d_x	The horizontal gradients
d_y	The vertical gradients
$\mathbb{D}$	The difference map after thresholding
e	Event
$\mathbb{E}$	A stream of events
$\mathbf{E}$	The event stream in a slice
F	Total number of frames
F_1, F_2, F_3	The vectors of higher order statistics
F_4	The maximum value obtained by RLE in a stream of events
F_5	The maximum timestamp in seconds in a stream of events
F_6	The number of ON events in a stream of events
F_7	The number of OFF events in a stream of events
f_t	The frame at time *t*
G_d	The number of ground truth answers
G_t	The gradients of the spectral entropy based temporal salience
H	The number of the temporal groups
I_k	The intensity of pixel k
$\mathbb{L}$	The length of the event stream in a slice
L	Total number of classes
l	Class label
M	The number of blocks in the saliency region
m_ℓ	The maximum events over the slice length
N	The total number of events

n	The index of the event
$\hat{P}$	The normalized power spectral density
P_d	The number of participants
P	The pixel in the chip's array
p	The polarity of an event
$\acute{p}$	The new polarity of an event
$\mathbb{R}$	The set of all RLE codes of a stream of events
R_t	The salience region
$\mathbb{S}$	The power spectral density
$S_{\hat{\mathcal{E}}}$	The spectral entropy based temporal salience map
S_ℓ	The total number of events in the 3D window-slice
s_g	The set of the events that are successive
s	The video sequence
t	The timestamp of an event
V	Total number of videos
V_d	The number of the video sequences for the dataset d
$\vec{v}_t$	The HOG-S feature vector of the frame t before the normalisation
$\hat{\mathbf{V}}$	Total normalised HOG-S feature vectors for a single video
$\hat{v}_t$	The normalised HOG-S feature vector of the frame t
$\tilde{v}_t$	The final feature vector of the frame t
u, v	The frequency coordinates
W	The number of the time intervals
x, y	The spatial coordinates
z	The third dimension of the events mask

List of abbreviations

AAL	Ambient assisted living
ADL	Activities of daily living
AE	Ambient environment
AER	Addresses-event representation
ADL	Activities of daily living
ANN	Artificial neural networks
APS	Active pixel sensor
ATC	Asynchronous temporal contrast
AUC	Area Under Curve
BoW	Bag of Word
CNN	Convolution neural network
CRF	Conditional random field
DBN	Dynamic Bayesian networks
DSP	Digital signal processing
DVS	Dynamic vision sensor
ECG	Electrocardiogram
FFT	Fast Fourier transform
fps	Frame per second
GCNN	Graph Convolution Neural Network
GMM	Gaussian mixture model
GUI	Graphical user interface
HAR	Human action recognition
HBA	Human behaviour analysis
HRLEP	Histogram of run length encoding of polarities
HMM	Hidden Markov model
HOG	Histogram of oriented gradient

HOG-S	Histogram of oriented gradient of salience
HOF	Histogram of Flow
HVS	Human vision system
IADL	Instrumental activities of daily living
IPs	Interest points
KNN	k-Nearest Neighbour
LIF	Leaky integrated and fire neurons
MBH	Motion boundary histogram
MEI	Motion energy image
MEFs	Motion event features
MHI	Motion history image
MMCC	magnitude of mean cross correlation
NVS	Neuromorphic vision sensing
PCA	Principle component analysis
PIR	Passive infrared
RFID	Radio frequency identification
RLE	Run length encoding
RNN	Recurrent neural network
ROA	Region of action
ROC	Receiver Operating Characteristics
SE	Smart environment
SIFT	Scale-Invariant Feature Transform
SNN	Spike neural network
STIP	Space time interest point
SVM	Support vector machine
QSVM	Quadratic support vector machine

Declaration

This thesis is the result of my own work, ideas, experiments and has not previously been submitted or accepted for any degree other than Doctor of Philosophy of the University of Sheffield. This thesis includes materials, which have been appeared in conference papers and submitted into journal papers.

Acknowledgements

I would like to take this opportunity to express my gratitude to my supervisor Dr. Charith Abhayaratne who provided insight and expertise that greatly assisted the research. I am especially indebted to my supervisor for sharing his pearls of wisdom and experience during my doctoral studies.

Also, I am sincerely grateful to my family: my wife and my children, Ali, Hasanain, and Zainab, and my family in Iraq, my parents, brothers and sisters, my friends for their love and support in my PhD journey.

Finally, special thanks to the Ministry of Higher Education and Scientific Research (Iraq) for providing a scholarship for my PhD study.

Chapter 1

Introduction

1.1 Introduction

Today, the interaction between people and technology is considered the highlight feature of our daily life, especially with all the ubiquitous computing devices. However, it can raise concerns of interpolating the computation into the privacy of users by different aspects: such as embedding it in the technology itself or through using the computing devices, *e.g.*, computers, as everyday objects. When the three pillars, *i.e.*, human beings, computation, and the real world, of this interaction fuse in a seamless approach, a smart environment (SE) can be obtained [7].

SE is the environment that reflects the interaction between people and ubiquitous computing. SE can represent any area in which the needs of occupants are immediately served by acquiring and analysing the contextual information in order to enhance the condition of performing their usual daily actions/tasks [8]. Mainly, the communication and cooperation between the ambient devices and sensor networks in the environment itself seem to be the paradigm of SE. Accordingly, SE seems to focus on supporting the daily activities of persons in their environment regardless of its variation by enhancing their abilities in executing the tasks [9].

Typically, there are many scenarios to implement SEs in real life. Ambient assisted living (AAL) is considered a typical scenario that can include both smart health, which is concerned with the independent home care monitoring, and smart home paradigms [7]. AAL is considered a new area of applications that can assist in finding solutions for the raised longevity problems, aiming to target the ability of older adults to perform their activities of daily living (ADLs) in

assistive environments (AEs) [10].

The concept of ADLs, which is used to assess a person's functional condition to live independently, is addressed by exploiting the benefits of AAL technologies to safely help older adults perform their daily activities [11]. Therefore, it is extremely important to automatically monitor and evaluate a person's ADLs by distributing sensors in smart areas, such as homes. Nowadays, various types of sensors are widely developed and used in smart homes to provide relevant information by analyzing the data obtained [12]. Table 1.1 summarises commonly used smart home sensors to monitor ADLs.

In Table 1.1, we observe that there are variations of sensor technologies that can be used to monitor ADLs and outputs useful information to identify the actions. One of the most functional sensors in this application is the vision-based sensor. A video camera is a preferable sensor that is exploited by researchers to collect data in vision-based monitoring approaches [13]. Furthermore, the vision-based sensor is considered a rich resource of information because detailed data can be obtained [12]. This vision sensor can be installed to monitor and recognize actions based on the obtained data. This vision-based monitoring system witnesses a growth towards building automated monitoring systems, increasingly to set-up cameras rather than other sensors in in-home assisted systems [14, 15, 16, 17, 18].

However, vision-based in-home monitoring raises the concern of persons about violating privacy [14, 11], since the vision sensor outputs intensities-based data. These data are easy to violate, which restrict the usage of vision sensors in the home monitoring scenario.

Recently, the neuromorphic vision sensor, which is inspired by the human's eye retina, is presented [19, 20] and used in the application of the computer vision. This sensor outputs the

Table 1.1: Summarization of common sensors used in AAL environment.

Sensor	**Type of measurement**	**Task**
Infrared	Motion	Localization
Radio Frequency Identification (RFID)	Motion	Objects usage
Pressure	Pressure on	Fall detection
Magnetic switches	Door/Cabinet	Object usage
Ultrasonic	Motion	Localisation
Video camera	Activity	Localization
Microphone (Audio)	Activity	Object usage

change of the intensity instead of acquiring the intensity, providing a new vision domain that can be used in the applications of in-home monitoring. However, the output of this new sensor is different compared to the standard camera. Therefore, we need a new method that can explore the output of the neuromorphic sensor for semantic tasks, such as action recognition.

1.2 Motivation

To date, we have seen a significant trend in using the vision-based sensor for monitoring and recognition of the daily activities of human in the literature. In-home monitoring and their contribution to healthcare for the ageing application have emerged significantly according to the needs for designing smart monitoring and recognition systems that can assist the elderly persons to live independently and safely. However, this application offers several challenges that have risen with incorporating the vision sensor for monitoring people in their properties. The main concern is the privacy issue since the camera is an intrusive sensor that violates individuals' privacy. Generally, there are technical and non-technical solutions presented to tackle this problem reasoning to increase the acceptance of using the camera sensor in the assisted living environment. Though, even with these solutions, which have been contributed to addressing the privacy problem, the trade-off between the privacy preservation and the ability to explore the obfuscation domain data is still the main challenge that affects the practical usage of such systems in real scenarios [11].

In the same context, technical and technological solutions have been proposed in this thesis. However, the proposed solutions aim to provide a reliable and robust framework which maintains the identity, *i.e.*, privacy, and, at the same time, provide a strong anonymity domain that is reusable to recognise the actions regardless of the trade-off between the privacy and intelligibility.

Non-technically, the human possesses one of the most complex visual systems and has an extraordinary ability to respond to the most important visual details that attract the human visual system (HVS). Modelling the visual attention of HVS plays an essential role in understanding the visual information in the scene [21]. This mechanism of visual attention targets the salient areas instead of the entire scene. The potential usefulness that arises from modelling the vi-

sual attention is manifold and can be exploited in several applications, such as removing the redundancy in visual data [22]. Therefore, modelling of this visual attention temporally, *i.e.*, temporal salient regions, provides a rich source of information reflecting the behaviour of the action overtime on which useful information can be extracted, and efficient recognition can take place.

From the technical perspective, the new revolution camera, which is also inspired by the HVS, *i.e.*, neuromorphic vision sensor (NVS), offers a tool that acquires useful salient information from the scene avoiding the redundant visual data [20]. This sensor seems to address privacy since the output of the sensor does not include intensity information, instead the intensity changes over time are obtained. This device is data-driven which relies on the magnitude of the captured motion, avoiding recording the redundant intensity information [23]. This camera outputs a stream of intensity changes, *e.g.*, events, instead of the intensity values making is a useful candidate in the application of AAL.

This thesis, therefore, aims to provide technical and non-technical solutions for the privacy issue in visual data and targets the utility of obfuscated data for human action recognition. The main research question is to obtain a reliable anonymisation method and provide useful information for exploring the utility of the anonymity domain for human action recognition. The obfuscated data and their improved accuracy is achieved without considering the trade-off between privacy and intelligibility. The combination of privacy and utility can provide a reliable, robust and flexible framework beyond solely the privacy that uniquely includes privacy and utilities together without using distinct models.

1.3 Key contributions

The main contributions of this thesis are:

1. **The proposal of a new vision-based privacy preservation method based on the temporal salience detection for a useful anonymity domain**

   The proposed anonymisation method aims to make the vision sensor reliable and increase the acceptance of a video camera used in an assisted living environment. The proposed method focuses on modelling the action's silhouette instead of modelling the subject's

silhouette. This modelling creates different silhouettes for different actions based on the dynamics of each action over time. Moreover, the action content is preserved as well as the temporal redundancy in the consecutive frames is omitted. The obtained temporal salience map omits the intensities and they are substituted with the saliency values in the anonymity domain. The subjective evaluation explains the robustness of the proposed method in the context of privacy compared to the current work, which is included in **Chapter 3**. The results of this evaluation prove that maintaining and anonymising only the action's relevant human body parts to construct the salience map leads to a higher level of privacy.

2. **A new descriptor for feature extraction from the action model in the anonymity domain**

   The proposed descriptor aims to emphasise the utility of data in the anonymity domain, which are modelled using the proposed method in **Chapter 3**. This new descriptor targets the region of the salient and avoids extracting features from the representation of the redundant data in the map. This outputs a robust description that is used to recognise the action. The robustness of the descriptor refers to the method in which the action is modelled, which, in turn, improves the discrimination among the actions and increases the similarity inside the action itself. The experimental evaluations show that the containment of this descriptor reduces the error rate in the matching process and improves the accuracy level of DHA, KTH, UIUC1, UCF Sports, and HMDB51 datasets by 3.04%, 3.14%, 0.83%, 3.67%, and 16.71%, respectively.

3. **A new method for exploring the neuromorphic sensing domain to represent the actions by extracting thier local and global characteristics**

   Introducing a new concept in vision-based sensor technologies, *i.e.*, neuromorphic vision sensor, and the nature of the data obtained from this device, opens up a new research area for the exploitation of this device in the application of privacy preservation. This camera outputs a stream of events synchronously, which indicates the intensity changes at each location over time rather the intensities. The output of this sensor seems to be useful to provide a reliable and robust anonymity domain. **Chapter 5** of this thesis,

therefore, proposes this device to address the privacy issue. Furthermore, this chapter presents a new method to extract features by exploring the neuromorphic domain for daily human actions. This proposed descriptor in this chapter analyses the patterns of events locally and globally to capture the local and global representation. Then, these features are fused to construct a robust description vector to recognise the action. In practice, several experiments are conducted using challenges datasets, and we demonstrate the ability of the neuromorphic domain to provide useful information that is exploited to provide a robust descriptor. The evaluation stage demonstrates the outperforming of the proposed method compared to the existing work on exploring the neuromorphic domain for HAR and the improvement of the accuracy rate by 0.54%, 19.42% and 25.61% for E-KTH, E-UCF11 and E-HMDB50 datasets, respectively.

1.4 Publications

Part of the work in this thesis has been published in the following conference papers:

1. S. Al-Obaidi and C. Abhayaratne, "Privacy protected recognition of activities of daily living in video," in *3rd IET International Conference on Technologies for Active and Assisted Living (TechAAL 2019)*, 2019, pp. 1–6.

2. S. Al-Obaidi and C. Abhayaratne, "Temporal salience based human action recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2017–2021.

This work is also submitted in the following journal papers:

1. S. Al-Obaidi and C. Abhayaratne, "Making Sense of Neuromorphic Data for Human Action Recognition," submitted toward IEEE Access, March, 2020.

2. S. Al-Obaidi and C. Abhayaratne, "Privacy Aware Human Action Recognition," submitted toward IEEE Access, April, 2020.

1.5 Outline

The thesis contains six chapters, which are listed as follows:

Chapter 2 provides a detailed description of the existing work on vision-based ADLs in AAL. Section 2.2 gives a detailed description and the current contributions on using different types of vision sensors for monitoring the individuals in assisted living environments. Section 2.3 illustrates the existing anonymisation methods to preserve privacy based on the sensors introduced in the previous section by explaining the exploitation of the anonymity domain for action recognition. Then, the feature learning and deep learning methods for ADLs recognition are reviewed in Section 2.4. This is followed by reviewing the RGB-based learning trend in Section 2.5. In Section 2.6, the new neuromorphic sensing domain is reviewed regarding HAR. Finally, the details of nine datasets which are commonly used in computer vision applications are explained in Section 2.7.

Chapter 3 introduces a novel method for privacy preservation by modelling the action based on the estimation of the temporal saliency. Section 3.1 provides a general introduction of the proposed method. Section 3.2 discusses the related work on concealment of the identity and creating the anonymity domain. Section 3.3 presents the proposed temporal salience detection to address the issue of privacy and introducing a useful anonymity domain. Section 3.4 shows the performance evaluation and discussions in terms of anonymisation, followed by a summary of the work in Section 3.5.

Chapter 4 targets the utility of the outputs; *i.e.*, obtained anonymity domain, of the proposed temporal salience anonymisation presented in **Chapter 3**. This chapter proposes a new method to explore the obfuscated data to extract a robust descriptor for action recognition. Section 4.1 introduces the proposed temporal salience-based action recognition. Section 4.2 discusses the related contributions on exploited HAR in in-home monitoring systems. Then, the details of the proposed method to explore the temporal salience are explained in Section 4.3, followed by an evaluation of the utility of the proposed method based on conducting several experiments using different datasets on different classifiers in Section 4.4. Finally, Section 4.5 summarises the work and its performance.

Chapter 5 presents a new domain to the issue of privacy and a new method to explore

it. In this chapter, we see how this new vision sensor is adopted for action recognition by the intelligibility of the outputs of this device. Section 5.1 introduces the differences between the neuromorphic vision sensor and the standard vision sensor. Section 5.2 reviews the related work on exploring the neuromorphic domain and the challenges that have not been addressed yet. The principle of using this sensor is explained in Section 5.3 followed by detailing the proposed method to explore the output domain of the neuromorphic vision sensing globally and locally in Section 5.4. The performance evaluation of the new descriptors on this new anonymisation domain is shown in Section 5.5. Finally, this chapter is summarised in Section 5.6.

Chapter 6 concludes the contents of the thesis and outlines the future directions of the ambient action recognition studies.

Chapter 2

State of the art

2.1 Introduction

In this chapter, we review the most relevant existing work of video-based anonymisation in the field of ambient assisted living (AAL) and its dominant area, *i.e.*, HAR. In general, the anonymity, *i.e.*, privacy preservation, can be seen from two perspectives: technological and technical. Technologically, in the in-home monitoring, the selected vision-based device can provide the required anonymisation to conceal the identity without the need to conduct any image or signal filtering algorithms to satisfy the anonymity. From the technical perspective, instead of using specific camera technologies, the standard RGB camera is used to acquire video sequence and applies a filtering or masking algorithm, for instance, to cover the identity.

This thesis aims to develop a reliable and useful anonymity domain to address the issue of privacy and explore the anonymised data for HAR regardless of the trade-off between the privacy and utility. To comply with these problems, we divide this chapter into six sections. Initially, we review the type of visual sensors that are used to address the privacy issue in AAL. Then, the existing methods to obfuscate the vision-based privacy are classified into two categories explaining the pros and cons of each category and the usefulness of their anonymity domain, followed by reviewing the methods of recognising the vision-based daily human actions. The new vision domain, *i.e.*, neuromorphic domain, is reviewed later. Finally, we give information about the datasets that are used in this thesis. In the following sections, the headlines mentioned above will be explained in separate corresponding sections.

2.2 Visual sensors in assisted living

In recent years, the designing of automated in-home monitoring systems has risen according to the desire of most elderly people to independently age in their home. Although such ageing-in-place systems seem to offer a promising service for higher senior health care, their success depends on improving the technologies of AAL. AAL provides tools for monitoring the ADL to improve health conditions of older adults life.

Usage of the sensors for monitoring and assessment of ADL is considered a crucial element for AAL. The sensors impose several technical and non-technical issues, such as the choice of the sensor, ease of use and installation, which are required to be considered for successful and useful ADL monitoring. Nowadays, various types of sensors are developed and utilized in smart homes to provide a significant source of information [12]. In general, sensors are categorised into wearable and non-wearable according to the localization of the sensor [11]. On the one hand, the wearable sensors, such as body-worn devices, are attached to a person or their clothes to measure the vital health metrics and motion characteristics. On the other hand, non-wearable sensors are localized in stationary places of the environment and mainly used to detect the subjects.

A video camera, for example, is considered a preferable sensor that is exploited by researchers to collect data in vision-based monitoring approaches [13]. Furthermore, it is considered a rich resource to obtain detailed information [12]. This vision sensor can be installed to monitor and recognize activities based on the obtained data. In some applications of AAL, such as, fall monitoring and detection, the trend of using intelligent vision-based systems is highly demanded. It witnesses a growth towards building automated monitoring systems to set up cameras rather than other sensors in in-home assistive systems [15]. In the following, four candidates vision-based sensors, e.g. low resolution, high definition, RGB-D and thermal, in the application of AAL, are reviewed explaining the advantages and disadvantages of each one of them.

2.2.1 Low resolution vision sensor

The low-resolution visual sensor is used to reduce the concern of persons with respect to privacy issue since this device captures low-resolution visual information. In general, this sensor is restricted to the applications of behaviour monitoring, such as fall detection and sleeping.

Culurciello *et al.*[24] investigated an asynchronous temporal contrast (ATC) vision sensor to detect motion. The ATC sensor responds to the changes in the light intensity of pixels to raise alarm of an accident of the fall in the elderly care environment. The imaging system is placed in front of the subject with a distance of 3 metres and a height of 0.8 meters to capture an array of pixels with resolution 64×64 pixels. In this sensing work, multiple side-view images of the scene are taken, alerting that an accident has occurred. This sensor is characterised by its ability to instantaneously produce a motion vector to be computed and forwarded in order to report accidents. This sensor is non-intrusive, has low power consumption, is small in size, and is a self-contained system. Moreover, the most important challenge that it can overcome is the privacy violation problem because the details of a patient's appearance are filtered out and processed locally.

Another low-resolution sensor technology is presented by Nyan *et al.*[25]. This sensor comes with a lower resolution, which is 30×30 pixels, compared to ATC, and is used to build a system for monitoring mobility patterns of older patients with cognitive disorders. A network of these low-resolution cameras is proposed to robustly track people based on distributed processing architecture. Despite the fact that multiple visual sensors are installed, the architecture is considered a much cheaper solution than other solutions in terms of using high-resolution cameras. Using these sensors can reduce cost, computational requirement, power consumption, and privacy concern. In addition to detecting motion, this system can recognize between movement of people and non-human objects. However, acquired images contain artifacts due to electrical interference as well as significant shading can result. Thus, image pre-processing techniques, such as, devignetting, are required to remove the effect of shading in order to correct the low brightness of images.

These sensors are used again in [26] and [27] to analyse the sleep patterns and behaviour of elderly persons, respectively. A network of 10 low-resolution sensors are used to analysis long-term behaviour (10 months of senior citizen activities), such as, eating, cooking, sleeping,

etc. The motivation is how to use low-cost sensors with little privacy concern. At the beginning, because the quality of captured images is poor, de-noising and sharpening techniques are used to remove noise and improve the quality of the images. Then, a foreground subtraction technique is applied to track the motion patterns of observers. Although the result is reliable, using multiple cameras increases the complexity of computation.

Despite the solution that this sensor offers to address the issue of privacy, the quality of the captured visual data is poor which affects the usefulness of the features that can be extracted from this low-quality data. Accordingly, the low-resolution camera-based system requires algorithms to enhance the quality of the captured visual data, which means extra computations.

2.2.2 High definition video camera

There are attempts to use the video camera to develop intelligent video surveillance systems for analysing ADLs of elderly people to offer a safe environment. Thus, different frameworks are recently proposed to use a high definition video camera for indoor monitoring. The video camera is classified as a sensor for AAL to monitor activities in smart homes [28].

To this end, Foroughi *et al.*[29] used a video camera with resolution 320×420 pixels for monitoring movement pattern to detect falls. A fixed video camera is used to record 50 video clips, representing 10 kinds of activities, which were captured and recorded in AVI format at 30 frames per second. Motion information, motion-energy image and motion-history image, is extracted from the video sequence to reduce both the space and the time consuming. Then, these features are fed directly to a classifier or firstly reduced by using eigenspace technique and then transferred to the classifier. However, since the video sequences are recorded in a lab environment, the results might be different in the real world.

The claim that the vision sensors can collect data which is difficult to gain by other sensors is presented by Skubic *et al.*[30]. Therefore, the proposed framework is to use a network of video cameras for monitoring elderly adults in their home to detect falls. The network contains six stationary video sensors, which are installed in the room with motion sensors to work together to discover the presence of activity by persons and to distinguish between visitors and residents. The privacy issue is addressed by applying a background/foreground subtraction approach to extract the silhouettes of the person. Because there are several cameras that are installed with

numerous sensors, moving to the real homes can be considered a challenge with respect to the amount of equipments and effort to install all these sensors in one place.

Building a real-time fall detection system based on vision sensors is considered a challenge in real life. Accordingly, a single high-speed camera with resolution 1280×1025 pixels is used by Shi *et al.*[31] to record the falling motions. It is positioned orthogonally with the direction of the fall and 5m away from the subject. Each sequence of recorded frames is analysed to compute the change in the inclined angle with respect to the time in order to alarm the airbag system and make it operate. The obtained data is firstly transferred into a digital signal processing (DSP) system, which is developed for real-time fall detection before the airbag is opened. Fast Fourier transform (FFT) and binary SVM filter are incorporated into an embedded algorithm to discriminate regular motions, such as, walking, sitting, and standing, for example, from falling.

A static video camera is used in CIRDO project [32] for monitoring older adults to detect falls automatically. After capturing a video stream, the position and the situation of the person are continuously extracted from the processed stream to detect falls. The video is processed in an automatic way to maintain the privacy of persons. Accordingly, it is captured and processed directly without recoding a copy and sent to the server to avoid operator intervention. In the stage of video processing, the silhouette is extracted based on foreground/background extraction, which is used to obtain useful characteristics for tracking the movement to identify the fall event. The scenes may suffer from the problems of changing light conditions and modifications of background. In this context, a Gaussian mixture method is used to enhance the illumination in both chrominance-luminance and chrominance-only spaces and generate a global illumination. Using one static camera seems to limit the field of view and restrict the camera range. However, this problem is solved by installing audio sensors in the region outside the camera range.

To avoid the field of view limitation, an omnidirectional camera can be used for monitoring and tracking because it can cover a large vision field. Thus, in [33], two omnidirectional video cameras are used, one is mounted on the ceiling and the other positioned on the sidewall, for monitoring the room space to introduce information about the actions of elderly people. They are captured video sequences with resolution 640×480 at 8 fps. The captured visual data is pre-processed by Gaussian mixture model (GMM) techniques to extract the foreground object.

In addition, artifacts are processed by removing small changes in the chromaticity and intensity. However, long time occlusions caused by multiple person participation in the same place can lead to a high missing rate.

2.2.3 RGB-D camera (*e.g.*, Kinect)

The entertainment device, Kinect, seems to offer a promising trend toward enabling detection of the human body and its movement without having to put any markers on it. It has become a preferred device to set-up in smart environmental systems following its capability and cheaper price [34]. Three sensors, RGB camera, IR projector, and IR camera, are integrated in Kinect to capture human body motions. It can preserve the privacy by acquiring depth images [35]. In elderly care, Kinect and its application can mainly address the need for alarm tools in in-home monitoring fall detection and risk reduction systems [36].

Falling is considered the indirect reason of death in the elderly. Several automatic approaches are proposed to monitor senior citizens to measure the expectation of a fall before it happens based on Kinect. In this context, Kinect device is used by Parajuli [37] to obtain 3D information which is recorded by 3D depth sensors. The data is collected in frames, which represent the posture of the person with its twenty joints (Hip Centre, Spine, Shoulder Centre, Head, Left Shoulder, Left Elbow, Left Wrist, Left Hand, Right Shoulder, Right Elbow, Right Wrist, Right Hand, Left Hip, Left Knee, Left Ankle, Left Foot, Right Hip, Right Knee, Right Ankle and Right Foot). Finally, a high dimensional feature vector that concatenates the coordinates of joints has resulted. It is positioned at the height of 1 metre from the ground. The obtained data is previously processed with data transformation and cleaned to adjust the distortion and reduce the scale of data to decrease the complexity, respectively. However, changing the scale of data can lead to the inaccurate performance of the classifier.

Kinect is exploited by Booranrom *et al.*[34] to assist older people to be secure and live independently. In particular, it is applied to a system to help older adults to switch on/off electrical devices without interacting them as well as raising alarm when elderly persons fall out of bed and any irregular behaviour during sleep. It can maintain the privacy of persons in both daytime and nighttime because a human skeleton is only detected and not any image. However, it may fail to detect the skeleton when the human body is covered by a blanket, for

example. Furthermore, the ability to detect all the human skeleton joints is inaccurate because only one Kinect camera is installed.

Moreover, Kinect is applied in an automatic system for monitoring the exercises that are done by elderly persons and to obtain on-line feedback [38]. A single depth camera positioned at about 3m distance from the subject is used to capture the whole body and estimate the poses. It is incorporated with nine infrared cameras, which are fashioned circularly to cover workspace. All these sensors are aimed to capture human poses and record 3D marker data. Although Kinect generates less precise 3D data than infrared cameras, it is preferable because it is unnecessary to do any user interaction, calibration or correction and therefore can preserve privacy. However, frequent errors in pose tracking can result due to the use of a single Kinect camera. Occlusions, such as self-occlusion by other body parts, or clutters seem to lead to frequent fails.

In an indoor environment with low conditions of light, Kinect may provide a solution to capture images and detect subjects because it has both infrared laser-based IR emitter and an infrared camera. Therefore, it is used by Michal and Bogdan [39] towards reliable, unobtrusive fall detection of elderly adults. However, its usage is restricted to be set-up in dark rooms. After capturing a depth image, the foreground object is extracted through the differencing of the current image from stored reference depth images without a subject.

The majority of Kinect applications for monitoring elderly people are restricted to the fall detection issue. The reason can be regarded with the risk of fall injuries for elderly adults and the physical and psychological problems behind it. However, the occlusions in the monitored scene can affect the accuracy of Kinect.

Following the limitation of the field of view, probably, integrating other vision sensors, such as video camera, with Kinect is valuable to track persons in out of range because Kinect is statically set-up. In [40], Kinect is integrated with a single RGB video camera to monitor adults for fall detection. Therefore, to meet the requirement of privacy concern, a foreground/background subtraction approach is applied on RGB camera video to extract foreground map which is used for human tracking rather than camera image. In the same context, Planinc and Kampel [41] integrated depth camera (*e.g.*, Kinect) is used with a 2D camera to build an automatic system for fall detection as well. Here, firstly, motion detection and background subtraction approaches are applied to extract the persons to enable detecting fall solely.

2.2.4 Infrared or thermal camera

Another imagery sensor that is exploited for monitoring elderly citizens is the infrared camera. It can serve to maintain the privacy of the observer. Since infrared cameras are less sensitive to brightness and low light conditions, Tzeng *et al.*[42] used an infrared camera for fall detection of elderly people in indoor environments. When the camera starts to track a subject, a sequence of images is recorded and then filtered and expanded to produce clean images and extract the features diagnosed by the expert system to alert fall detection. In spite of robust results, however, the infrared camera is unable to detect all the patterns of falls because it is difficult to identify the fall of older adults based solely on visual data. Therefore, another sensor, such as the pressure sensor, is combined with an infrared camera and an expert system addresses fall detection based on the assessment of extracted features of both.

In the same context of fall detection, Sokolova *et al.*[43] build an infrared video-based fall monitoring and detection system for elderly people. An infrared camera is used to record a video with resolution 720×480 pixels for monitoring static and dynamic falls of a single person. After capturing images, each image is normalized to unify the scale of pixel values and then is subjected to a binarized thresholding process to isolate the elements of the human. This segmentation is performed on both single infrared image or a sequence of infrared video images to detect static or dynamic falls, respectively. Although the findings and the performance were better than other technologies, testing the proposed method in the real world with a single camera can change the expectations.

The thermal camera is also utilized to track people in ambient assisted living environments [44, 45]. We found two scenarios to use this type of sensor: ceiling camera [44] and standing camera [45]. In the first scenario, the camera is positioned on the ceiling of the apartment (kitchen, for example) at the height of 4.4m to cover approximately 6m^2 of the monitored area. A sequence of low-resolution images with 80×60 pixels is captured and fed into a set of multi-hypotheses Monte-Carlo particle filters. Firstly, these images are pre-processed by static background subtraction to eliminate image noise and to isolate the foreground object. Then, 8-neighbour connectedness for each pixel is applied to segment the resulting foreground image. In the second scenario, the camera is fixed on standing at the corner of the room to measure the temperature of the body using a low-resolution array of 8 pixels. The capture vector of pixels

is then sent through the recurrent neural network (RNN) to detect the falling and to launch the alert.

Recently, a low spatial resolution infrared sensor is used for ambient assisted living [46]. The proposed system is used to address privacy and daily action recognition. The thermal sensor itself retains the privacy since this sensor outputs 8×8 array includes the temperature as the pixel intensity. The obtained thermal images are fed into a 3D convolution neural network for the action representation.

However, the limitation field of view of the camera imposes to be mounted in an unusually high position which can lead to capturing non-human heat sources caused by some local and global sources, such as, some appliances and sunshine, respectively.

2.2.5 The Challenges

Mainly, there are common challenges that can be found when reviewing vision-based sensors and need to be considered and addressed. These challenges can be classified into technical and non-technical which are listed below. These limitations can affect the performance of the overall system and may lead to failing.

Technical challenges:

1. The choice and set-up of sensors.
2. Signal processing techniques.
3. Machine learning algorithms.
4. Limitation of camera field of view and distance.
5. Occlusions.
6. Illumination changing.

Non-technical:

1. Ease of usage and installation.

2. Privacy violation.

3. Real life environment.

Shadow field or light-less environments, for example, can significantly degrade the performance of video camera-based home monitoring [36]. Therefore, several vision-based tracking systems combine a variety of vision-based sensors in a single system to overcomes these obstacles [47]. However, using different vision technologies means different multi-modal processing and this multi-modality increases the complexity of the system. Moreover, how many sensors will be adequate to offer rich data? What is the vision sensor technology that is better than others? This needs to be addressed as well. Table 2.1 summarizes the main differences between all visual sensors against different challenges.

We argue that extracting a useful abstract from the visual data is more effective to identify the action instead of using extra computations, such as a multi-modal processing.

2.3 Vision-based camera anonymisation

In this section, the existing work on privacy preservation is presented, explaining the challenges that are still not addressed yet. In this reviewing, both the technical and non-technical solutions are discussed and argued, highlighting the weak points in these solutions that can be contributed to improvements in the performance. The issue of privacy is still the main challenge in the applications of healthcare and smart homes whenever the identity, appearance and face details

Table 2.1: Summarization of camera-based sensors capabilities against different challenges.

Challenge	Low resolution	High definition	Depth	Infrared
Light change	Sensitive	Sensitive	Less sensitive	Less sensitive
Cost	Low	High	Low	Low
Image processing	Required	Required	Required	Required
Privacy maintenance	High	Low	High	High
Field of view	Limited	Limited	Limited	Limited
Occlusion	Sensitive	Sensitive	High sensitive	Sensitive
Image quality	Low	High	Low	Low

raise the concern of the subjects. Privacy is also considered one of the highest demanded concerns in the Internet of things (IoT) systems [48, 49]. This challenge is compounded when the sensor is a camera which is one of the sensors that arouse the sensitivity of in-home monitoring applications. So, the researchers work hard on this challenge to meet the requirements of the subject to accept using the camera sensor. Because of using a video camera in our research, the following review focuses on the technical and non-technical vision-based solutions to address privacy in the anonymity domain and the existing challenges in these solutions.

2.3.1 Image processing-based anonymisation

Technically, several solutions based on processing visual information have been presented. Most of these methods leveraged from image processing techniques to cancel the people's sensitive vision-based information. Accordingly, blocking [50], cartooning [51], and masking [52], to obfuscate the sensitive information are the most dominant approaches used in this domain.

Blurring is one of the standard algorithms to protect people's privacy through blurring the face in the vision-based monitoring systems [53, 54, 55, 56, 57]. Though blurring the entire person's body is more useful to provide a full identity obfuscation [58, 50] since other body parts can reflect the personality profile. This filter modifies each pixel in the privacy region or the whole image by applying the Gaussian function using the neighbouring pixels.

Pixelating is another widely used method adopted to obfuscate the privacy in vision-based assistive living environments [59, 60]. According to this filter, the image/frame is partitioned into blocks of pixels, and the average value of the intensities in each block is assigned into those pixels in the block. This method reduces the resolution of the privacy region and provides strong data abstraction. However, this high-level of privacy leads to losing a considerable amount of visual, leading to a low-level of utility [61].

Blocking/masking is a way to preserve the sensitive regions in still images and video sequences by covering the identification/sensitive appearance information to obfuscate the privacy. This approach takes several patterns; such as, blocking with a grey-level block [50], background or black bounding box [62], or a solid silhouette [63]. Masking or blocking can be in two different themes: face masking [52] or the whole body covering [63].

Encryption-based privacy obfuscation algorithms are used to conceal the personal de-

tails in image/video [64, 65, 66, 67, 68]. The encryption is applied over the whole frame or the region of the sensitive data to achieve privacy. The encryption-based methods scramble the sensitive visual information to achieve a high level of privacy which means that the protected region of privacy is unusable without decryption of the encrypted data [69]. Therefore, to make the encrypted data intelligible, the original data has to be recovered, and this makes the privacy vulnerable to piracy and increases the risk of it being compromised, which causes people's concern. Besides, the encryption-based method is computationally expensive. Thus this privacy approach generally is not efficient for real real-time requirement vision-based surveillance applications [70].

Person removal/replacement are also used to address privacy in vision-based home monitoring. The methods that belong to this category include the techniques of cartooning [61, 71], 3D avatar replacement [72], cartoon picture replacement [73] and person removal [74]. The algorithms of cartooning, avatar and picture replace the real sensitive identity region by another representation of virtual reality. At the same time, person removal methods delete the sensitive information and fill the gap by image in-painting algorithms [75, 76].

The above mentioned vision-based privacy preservation methods provide a low-level of utility since increasing the level of privacy impairs the intelligibility of the obfuscated data [77]. This challenge is caused by considering the trade-off between the privacy and utility of the anonymised sequences for monitoring tasks [78], and this trade-off affects the reliability of these methods. However, other methods lose this balance, such as encryption-based anonymisation since the encrypted information can not be used without a decryption method. In general, a high level of privacy protection leads to low-level of utility and vice versa. This trade-off between privacy and utility is one of the significant challenges associated with using the vision-based sensor in AAL. Furthermore, most of these filters are weak to protect privacy and fail against attacks [79]. The reason is that these filters are still keeping the intensity values in the obfuscated version; therefore, it is easy to recover the sensitive information.

2.3.2 Sensor-based anonymisation

There have been a few solutions to address the privacy concerns that the sensor as a device can offer. However, mainly there are two cameras: low-resolution sensor [80, 27, 81] and neuro-

morphic sensor [19] that can create data in the anonymity domain without the need to conduct additional data processing. The use of low-resolution sensors adopts a network of extremely low-resolution cameras [27, 80, 82, 46] or low-resolution colour sensors [81] to capture low-resolution visual images. This sensor has been successfully retaining privacy in the in-home activity recognition systems [80]. Low-resolution visual information has been explored for behaviour understanding [27] and object localisation [81]. Recently, this sensor is explored to recognise multi-class human action systems. In this context, Tao *et al.*[82] exploit a set of three low-resolution cameras to recognise human activities based on hand-crafted features learning. This topology is developed in [46] by exploiting the 3D CNN to leverage from the deep leaning leading to increase the accuracy of recognition by 10%. However, these sensors are more sensitive to the changes in the light conditions [27, 81], resulting in less accuracy in activity recognition.

The Kinect is another vision sensor that can be used in an assisted living environment to preserve privacy. However, the main obstacle is that the subject must be in the range of three metres in front of the sensor in order to acquire the visual information; otherwise, the Kinect fails to record this data.

Recently, neuromorphic vision sensing (NVS) camera has emerged as a new technology in the field of vision-based sensors inspired by the eye retina that relies on capturing changes in log intensities instead of recording colour intensities of the pixels[19, 20]. This sensor will be explained in details in Section 2.6 and Chapter 5, respectively. Because the NVS camera outputs a stream of events instead of frame-based visual information, appearance details cannot be detected or recognized. This concealment technique makes the NVS camera a suitable candidate sensor in the home monitoring applications to preserve privacy. To date, though the NVS camera has been used for HAR, exploring this sensor in the applications of assisted living has not been addressed yet. The domain of the data generated by this sensor inspired us to suggest the NVS camera as a new tool that can be exploited in the applications of AAL.

2.3.3 Action recognition based on exploring the anonymity domain

Exploring the anonymity domain to recognise the human actions acquired by the vision-based sensor is one of the main goals in the AAL system. This is important to track and analyse

the daily actions of the older people in order to alert the emergency in the case of an urgent situation. However, recognising the daily human actions using the vision-based sensor (or any sensor) must take into account the issue of privacy. Therefore, in-home monitoring/tracking systems must consider the concern of people, mainly whenever a vision camera is used for monitoring.

The trade-off between privacy and intelligibility is more important to satisfy the reliability and confidentially in the assisted living application. In this context, several researches have been presented to address the trade-off between privacy preservation and utility [83, 63, 25, 26, 27, 46]. These systems depend on feature learning or deep learning to explore the anonymity domain and recognise the actions. Although these methods have successfully achieved the utility, the accuracy needs to be improved without affecting the privacy since these systems are concerning about satisfying the trade-off between the privacy and intelligibility. In general, most of the filtering algorithms, which are mentioned in Section 2.3.1, decrease the recognition accuracy [83]. Therefore, it is essential to keep into account the level in which privacy is preserved completely.

2.4 The techniques of recognising the activities in assisted living (activities of daily living (ADLs))

In recent years, building automated in-home monitoring systems has been rising by the growing desire of older adults to age in their own home independently. Although such ageing-in-place systems seem to offer a promising service for senior health care, their success relies on the improvement of AAL technologies. AAL provides tools for monitoring and supervising the activities of daily living (ADL) in order to improve the living conditions of older adult life.

ADL can be defined as all the normal activities of everyday life that individuals can do in their daily living, such as dressing, eating, walking, bathing, sleeping, etc. Thus, the ability of a person to perform these activities can be used to assess the functional status of him/her to live independently.

Automated monitoring and assessment of the ADLs is a technological tool that can be used in AAL environments in order to assist older people in their homes [11]. In this context, sen-

sory technologies can be adopted for monitoring and acquiring information from the ambient environment (AE), then by feeding the gathered data into an automated system to recognize the ADLs. However, the type of sensor offers additional challenges that need to be addressed in monitoring ADLs. The sensors seem to impose several considerations, such as ease of use and configuration effort [84], cost-effectiveness and preserving the privacy of the user [85], that emerge and need to be addressed as well.

Generally, the sensors deployed to acquire data in the research of ADL recognition can be distinguished into two categories: wearable sensors and non-wearable (ambient) sensors. In the experiments with wearable sensors, the sensor is attached to the subject directly to measure and monitor location, blood pressure, pulse rate, and other vital signs. The accelerometers, gyroscopic, bracelet sensors, for example, are examples of wearable sensors. Although this type of sensor is accurate to localise the subject [11], many disadvantages can be found [27], such as the limitation of the battery life, missing data, and usage of multiple sensors attached to specific body parts to obtain more reliable measurements [86].

In the ambient sensors applications, the sensor can be mounted on the wall or embedded in the furniture. Broadly, the passive infrared (PIR), radio frequency identification (RFID), and camera sensors are considered widely used sensors in the research of smart environments [12]. However, the main drawback is that these sensors are less intrusive, especially vision-based sensors [86], which can raise the privacy concern of observers.

To recognise the activity of the observer based on the data acquired by the sensors, reasoning over the data is performed to detect the activities. There are two approaches to recognise human activities, *i.e.*, data-driven and knowledge-driven approaches, based on the sensor technology that is used to monitor the older observers [7]. In terms of ADL, the majority of research relies on a data-driven approach. This is the commonly used approach in this context. It requires data labelling, pre-processing, and training using machine learning techniques. Machine learning algorithms are considered the most intensive techniques that are used to recognise human activities based on the information feed from sensors [87]. Machine learning techniques have proven to strongly contribute to building systems to recognise the activities of human-based on sensory data [7]. These methods are able to manage the uncertainty and temporal information [88]. They are broadly classified into two categories: generative and discriminative model as

presented in the following sections.

2.4.1 Generative machine learning techniques

Generative models predict the most likely class based on estimation of the joint probability distribution of the samples and the labels [11]. There are several models that belong to this category. Typically, the hidden Markov model (HMM) is the most popular technique used in recognising the ADLs. Patterson *et al.*[89] used the HMM to recognise the routine morning activities based on data from wearable RFID, which is tagged on the subjects. Whereas HMM is used to recognise the hand-related activities, such as sawing, hammering, filing, etc., in [90] using data of two accelerometers sensors. Modayil *et al.*[91] exploited this model to discriminate between three ADLs, *e.g.*, drinking a glass of water, making a stir-fry and making jelly. The information to HMM is fed from a wrist-worn RFID reader and set of tags placed on objects. Other proposed works to recognise the daily activities are introduced in [92]. Other wearable sensors, such as smartphones, are exploited to recognise human activities by HMM. In this context, Nickel *et al.*[93] applied HMM to classify the biometric gaits of users using the accelerometer sensor in a Motorola smartphone. Other contributions based on using smartphones capabilities and HMM are found in [94] and [95]. The HMM model is also exploited to deal with a few action samples [96]. In this context, the information of motion is captured in 2D modelling, spatial and temporal, from sensory data of multicamera video sequences.

The second generative model that has been proposed in terms of daily human activity recognition is the Naîve Bayes classifier. This model is exploited in [97] to build an on-line daily monitoring system of cardiovascular disease patients based on the sensory data of both surveillance of electrocardiogram (ECG) and blood-pressure sensors. It is used in well-being applications by Shoaib *et al.*[98]. The sensory data of accelerometer and gyroscope, which are build-in smartphones, captures the information about the activities of observers. More recently, Wang *et al.*[99] applied the Naîve Bayes model to recognize the simultaneous and separated performed human physical activities using two smartphone sensors, *i.e.*, triaxial accelerometer and gyroscope.

Another model in this category is Dynamic Bayesian network (DBN). DBN is applied in [100] and [101] to monitor the activities and physical and cognitive capabilities of residents

in the smart home. However, the main drawback of the generative machine learning models is their limited ability to deal with varied sensory data [88] because they are considered static models [102]. Moreover, these models require that the training stage seems to be considered in order to recognize all of the possible observation sequences of an activity [88].

2.4.2 Discriminative machine learning techniques

Contrary, the discriminative models are the alternative approach in recognition of human actions. The efficiency of these models is potentially inherent in their ability to manipulate with high dimensional data [11]. There are a variety of statistical models that belong to this category. There are four popular discriminative models, support vector machine (SVM), conditional random field (CRF), artificial neural networks (ANNs), and k-Nearest Neighbour (KNN), popularly used by the existing work in the domain of ADL recognition.

SVM is considered the most widely used classifier in the domain of ADL classification. In this framework, Brdiczka *et al.*in [103] used this discriminative data-driven machine learning technique to label the training data by learning from data of the daily activities of observers. More recently, SVM is combined with generative models, such as Bayesian networks [104], in order to gain a more accurate hybrid model to recognize the activities of the users.

CRFs are considered flexible classifiers that can discriminate complex patterns of activities, *e.g.*, unarranged or unordered, [88]. CRFs are graphical models such as HMMs, but there are considerable differences between both algorithms, such as concurrent and interleaved activities are allowed with CRF. CRFs are applied starting from the simple scenario, such as in [105], to more enhanced versions of CRFs in [106] and [107].

The third machine learning technique that has been applied to recognise human actions in smart environments is ANN. This intelligent structure simulates the methodology of the biological nervous system to works and processes information in reality. A neural network is represented in the form of interconnected neurons, which are the processing elements of the network that can solve the problem. ANN has been used by Khan *et al.*[108] to recognise a set of daily physical activities ranging from static to dynamic activities.

Lastly, the KNN algorithm is a simple data-driven approach and used to recognise the human physical actions [97, 87]. This classifier calculates the distance between the input feature vector

and all references. The shorter distance is, the more similar to the compared action reference.

2.4.3 Vision-based deep learning for ADLs

Recently, the trend of using deep learning for vision-based AAL has witnessed an increased interest from the researchers [109, 110, 111, 112]. This emerging interest is due to several reasons; such as the high level of accuracy, no pre-processing, and no need to design a hand-crafting feature extraction, that the deep learning algorithms introduce. The deep learning is used to recognise a specific serious behaviour, *e.g.*, fall detection [109, 111] or normal daily activities [112]. However, these tools require a huge number of data for training and testing.

2.5 RGB vision-based HAR representation

In this section, we briefly present the recent work on HAR based on both local and global representation explaining the challenges that have not been addressed so far in each theme.

2.5.1 Global descriptors

The previous RGB-camera approaches using the global descriptors are primarily based on either space-time descriptors [113, 114, 115, 116, 117, 118, 119, 120, 121], shape-based analysis [122, 123], or deep learning based [124, 125, 126]. They either represent the action using the whole silhouettes or extracting holistic description from these silhouettes. In the first category of global description, the shape is encoded using a space-time representation along both the spatial and time dimensions to model the human motion. Different approaches, such as, the motion energy image (MEI) and the motion history image (MHI) [113] or volume [116], space-time shape [118, 114, 115], silhouette and optical flow [119], shape and flow [120], and silhouette [121] are applied to describe the actions. In shape-based learning, the global features; such as, appearance and motion features [122], frequency domain-based features [127] and hand gesture recognition [123], are extracted based on generating the silhouette or the shape of the human body.

These descriptors are sensitive to occlusion and cluttering [127, 128] leading to shortcom-

ings in the results [129]. In addition, these methods rely on using complex approaches, such as background subtraction, to extract and describe the foreground object. All these methods extract the whole body part without focusing on the most significant points in the body, i.e. the moving parts. Recently, the video-based saliency estimation offers the mechanism to determine the most dynamic parts of the human body based on plausible simulation of the human visual system. The calculated saliency map provides an efficient approach to highlight the salient parts of the body without using high complexity motion estimation algorithms and reducing the amount of the redundancy. Creating a temporal salience-based silhouette seems to provide a means to filter the redundant temporal information between the consecutive frames and highlighting the locations when the intensity changes are found. This can enable the descriptor to extract meaningful features that can accurately discriminate among the action.

2.5.2 Local descriptors

The previous works on extracting local features using vision sensors are primarily based on either detecting interest points (IPs) [4, 130, 131, 132, 133, 134, 135, 136, 137, 138, 139] or densely sampling [130, 140, 126]. These descriptors extract either the dense trajectories features or space-time interest point (STIP) based features. These features are common and more resistant against the background noise and the occlusions [141]. The local descriptors, such as HOG and HOF, are widely used to obtain the local features in HAR applications.

In IPs representation, the feature vector is extracted around the candidate points based on STIPs. The histogram of local features [4], SIFT descriptor [131], dense features [132], dense trajectories [133] are commonly used to explore these STIPs.

In the second category of the local description, the dense descriptors, *e.g.*, HOG, HOF, and motion boundary histogram (MBH), are used to encode the action exploring the whole image/frame regions. By applying these descriptors directly on the visual data, redundant information is used to describe the action which is inefficient since irrelevant content is included to encode the human action. Therefore, the extracted features usually are subject to extra filtration, such as the quantisation using a bag of features or Fisher vector methods [142], to improve the description of the action. However, these descriptors are typically suffering from many shortcomings when dealing with long-term actions [126] and exploring high complexity algorithms

to estimate the motion. Therefore, proposing action modelling and description without using complex algorithms and exploring the representation without post-processing is considered urgent in HAR.

2.6 Neuromorphic sensing domain for HAR

In this section, we will present the techniques that are used to generate the event stream firstly and then we will explain the existing methods on exploring the neuromorphic data for action recognition. Technically, the hardware and the software approaches will be introduced in details to explain the pros and cons of each one. Though the rarity of the research in the field of HAR which is still in early stages [143, 144, 145, 146], we can recognise two themes of the research area: behaviour monitoring [147, 148, 20] and high-level semantic action recognition[146, 144, 143]. Accordingly, the current work that has been done on each category will be explained in detail.

2.6.1 Dynamic vision sensor (DVS) data acquisition techniques

There are two frameworks to capture neuromorphic sensing data: using neuromorphic cameras or emulators. Both techniques achieve effectively in the applications of applying the neuromorphic sensing data to solve real-life problems. However, there are some advantages and disadvantages relevant to each technique that need to be considered before using them. In the beginning, we will explain each technique giving examples and then make a comparison between them.

Neuromorphic sensors

The neuromorphic camera has undergone many changes in the last two decades. The first model of this type of sensor was developed in 1992 by Mahowald (and his assistants) as part of his PhD project [149]. This silicon retina based sensor produced the events using addresses-event representation (AER) protocol. However, several problems were found in this design; such as there is a miss-match between silicon-based pixels since this camera combines two different paradigms with large pixel cells making it difficult to use it practically in the real world. In the

next couple of years, efforts were modest and insufficient to discover the potential computations of this type of camera and address the shortcomings of this camera. Since 2003, the idea of the AER-based sensor has attracted the researchers to develop new versions that address the potential problems of the conventional active pixel sensor (APS).

Commercially, Table 2.2 depicts the widely used branded neuromorphic cameras that represent the revolution in the vision sensors nowadays and summarises their characteristics. Some of these models contain an APS sensor to record intensity-based frames if needed.

Simulators and emulators

Due to the rarity of neuromorphic cameras and the high cost of them, several simulators and emulators have been developed to meet the increasing demand for this sensor. Some of these simulators are publicly available now to conduct experiments on converting the RGB versions of sequences to neuromorphic event streams. In the following, we explain a list of these simulators:

- **VSBE** [155]: It is a vision sensor behavioural emulator that can be used to simulate the architecture and the function of any neuromorphic vision sensor. VSBE provides a high frame rate up to 128 frames per second. This emulator contains PS3-Eye camera, which is expensive, providing the higher frame rate using mono and colour OVGA modes. This camera provides a (320×240) resolution to be compatible with the installed frame rate. This emulator is then linked into one of the large library tools used to develop the event-based sensors; *i.e.*, *j***AER** [156]. This emulator can be used to evaluate the performance of several NVS-based sensors; for instance, DVS and cDVS.

Table 2.2: Common Neuromorphic cameras and the main characterizes of these sensor: The characterizes focuses on the features of the resolution, pixel's size, dynamic range and if there is an APS sensor or non.

Device	Resolution	Pixel's size	APS included	Dynamic range (dB)
DVS128 [150]	128×128	$40 \times 40\mu m^2$	No	120
ATIS [151]	304×240	$30 \times 30\mu m^2$	Yes	143
DAVIS240 [152]	240×180	$18.5 \times 18.5\mu m^2$	Yes	130
DVS-Gen2 [153]	640×480	$9 \times 9\mu m^2$	No	>80
CeleX-IV [154]	768×640	$18 \times 18\mu m^2$	No	-
DAVIS346redColor [6]	346×260	$18.5 \times 18.5\mu m^2$	Yes	120

- **DAVIS simulator** [157]: The simulator generates three types of data streams: the event stream, the intensity frame and the depth map. This simulator is used to evaluate the new products of dynamic and active-pixel vision sensors.

- **ViSim** [158]: It is used to assist creating monocular or stereo camera trajectories and synthesize related ground truth. The simulator provides a user-friendly interface while adding our simulation functionality behind. It contains a Design View and a collection of Parameter Options to configure camera calibration, distortion coefficients, stereo view, and frame rate and travel time.

- **ESIM** [159]: This simulator combined the rendering engine and the event simulator. ESIM generates the stream of events by using an adaptive sampling method.

- **pyDVS** [160]: This emulator contains two parts: a conventional digital camera to record frames-based information and a PC to convert these frames into a stream of events. Figure 2.1 displays the layout of this emulator.

- **PIX2NVS** [161]: the PIX2NVS is a software-based released emulator available for academic research usage. The main application of this emulator was to provide a cheap tool to convert a conventional of video sequences into NVS streams of events. It is a software codebase used to generate neuromorphic vision streams from any pixel-domain video format. Conducting experiments using this emulator proved its suitability to generate the artificial NVS streams.

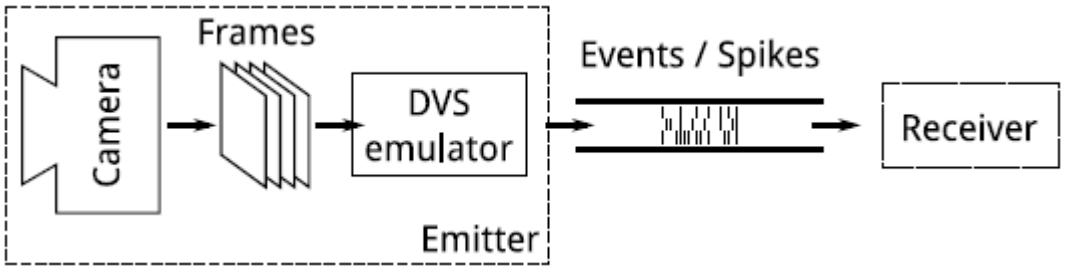


Figure 2.1: The layout if pyDVS emulator.

2.6.2 Behaviour monitoring exploring neuromorphic sensing domain

To date, the current work on using the NVS sensor has been limited to monitoring and analysis of specific human actions; such as finger/hand movement [147, 162] and fall detection [148]. If we start from the first research on this topic, we find that Belbachir *et al.*in [148] introduce a stereo vision system to monitor fall detection for older people. The system incorporates two DVS cameras to create a 3D spatio-temporal scene. From this volume, the depth information and the event rate are concatenated to detect the fall. This system offers promising results for real-time fall detection; however, the system fails to distinguish between the real and false rates due to use of a few features and considering a few examples of falls. The real-time applications regarding the usage of the DVS cameras attract the researcher to investigate algorithms for more interactive and immediate response applications. For instance, Lee *et al.*in [147] were firstly exploring the neuromorphic sensing data to present a real-time interface to recognise the moving finger that touches a mobile phone screen. The NVS camera has been used to track the finger movement and its direction by exploring the events, which are generated from the DVS camera, around the edge of the finger. To track the finger, a two-dimensional array of leaky integrated and fire neurons (LIF) are used while the direction is determined using winner-take-all configurations of neurons. Despite the outperforming of this proposal, its application is restricted to the mobile devices due to using a low-resolution NVS sensor which is 128×128.

In 2014 Lee at al. developed his first prototype for fingertip detection to hand gesture recognition [162]. The work presented a graphical user interface (GUI) control system by deploying stereo neuromorphic vision system. The spatiotemporal output of the stereo DVS cameras is fed into a spike neural network using LIF neurons to construct the feature vectors. However, developing the neuromorphic-based user interface for hand gesture recognition shows several challenges that degrade the accuracy of the recognition. This system was the first work that combined both the neuromorphic sensor and hand gesture recognition. To develop this system, Amir *et al.*[1] thought that this prototype could be improved by adding an event-based neuromorphic processor which includes a deep convolution neural network (CNN) to achieve a high level of accuracy of recognition of the hand gesture in the real-time. This processor had been already tested on recognising the playing cards [163], and the results inspired the authors to use this processor to improve event-based hand gesture recognition. The integrated processor runs

the CNN to process the stream of events passed from the DVS128 camera. Figure 2.2 shows the prototype of the system, including both the DVS128 camera and its attached processing unit. Again, the spike neural network (SNN) had been exploited to process the spikes or events to recognise the hand gesture by [164]. However, in [164], the events are fed to the classifier without any pre-processing. The prototype uses a spike neural network (SNN) with less layer compared to the system in [1].

Recently, both Wang *et al.*[165] as well as Chen *et al.*[166] made a different proposition to what is presented in the current work, focusing on exploring the event domain instead of accumulating the events into frame-based representation and learning. The prototype for hand gesture recognition presented in [165] had been tested on simple and short term datasets, and the results are not enough to judge on the superiority of this model compared to the existing methods in the field. Chen *et al.*in [166] avoided this obstacle by recording a larger scale hand gesture dataset with 2,040 instances of a set of 17 gesture classes. This model has a problem to deal with low hand gesture activities since these activities have low event density and confuse the classifier leading to reducing the accuracy of recognition. Table 2.3 summarises the characteristics of all the existing work on multi class-action recognition.

Figure 2.2: Hand gesture recognition system proposed by Amir *et al.*[1].

Table 2.3: Summarizing the main characteristics of event-based neuromorphic behaviour monitoring systems.

Feature	[148]	[147]	[162]	[1]	[164]	[165]	[166]
Application	Fall detection	Finger localization	Gesture	Gesture	Gesture	Gesture	Gesture
Real time	✓	✓	✓	✓	✓	✓	✓
Topology of camera	Stereo vision	Single	Stereo vision	Single	Single	Single	Single
Camera model	Two DVS128	DVS128	Two DVS128	DVS128	DVS128	DVS128	DVS128
Learning	Hand-crafted	Hand-crafted	SNN-based feature	Deep learning	Deep learning	Deep learning	Hand-crafted
Classifier	-	-	HMM	CNN	SNN	PointNet	RNN

2.6.3 High semantic multi-class action recognition exploring neuromorphic sensing domain

Exploring the NVS camera has been upgraded beyond a single action to highly semantic applications, including multiclass action recognition. Exploring the neuromorphic sensing data for HAR is still at the beginning. Most of the existing methods focus on converting the neuromorphic domain events into other domains; such as artificial frames [167], suitable to learn by the classifier or deep learning because the nature of NVS sensing data makes it challenging to use it directly. This approach of exploring NVS-based framing leads to loss of the high frame rate, which is one of the advantages of NVS domain over APS domain.

One of the early contributions in this category had been introduced by Sullivan *et al.*[146] that combines the neuromorphic camera with CNN. In the beginning, the motion is represented by extracting its direction and magnitude from the events and then feeding these features into CNN. The presented method relies on modelling the motion using the event stream. Accordingly, the events are converted into frame-based representation and then stacked into interval-based frames. The number of stacked frames each time depends on the level of the action, *i.e.*, fast action or slow action. Then, a new feature vector called motion event features (MEFs) is constructed from these stacked frames by counting and normalising the number of events in a grid centred at each location (x, y). Finally, CNN is applied to the obtained event feature maps to identify the actions. Baby *et al.*[143] follow the same idea to convert the events into frames/maps. However, the difference compared to [146] is that hand-crafted features are extracted from these constructed maps and used to train the SVM classifier instead of constructing

maps with CNN. The framework is started by converting the events from the neuromorphic domain into the intensity domain. By this, we think that the essence of the neuromorphic sensing data foregoes the events towards the pixels. This mapping approach aims to generate useful NVS frame representations that can be exploited by the conventional feature extraction method to extract a meaningful set of frame-based features. These extracted features are SURF and MBH, which are encoded using the histogram of the bag of features. These are combined and used to learn the SVM classifier. The results of recognition point out that incorporating the DVS for HAR is considered a promising framework in this field.

Recently, new emulators; such as pyDVS [160], PIX2NVS [161] and DAVIS simulator [157] have been designed and publicly available. These emulators provide cheap software-based tools to generate the events from the RGB video frames overcoming some of the limitations of the neuromorphic cameras and fill in the shortage of annotated NVS-based datasets. Using these emulators contributes to filling the gap of the lack of available labelled NVS training data. There is existing work exploiting these emulators for producing higher-level action datasets for HAR. Chadha *et al.*in [144] conduct experiments using PIX2NVS emulator. This software-based NVS sensor is used to generate large-scale labelled training datasets, *e.g.*, HMDB51. The polarities of the generated events are aggregated based on a specified time interval to provide frame-based representations suitable to learn the CNN classifier. The teacher pre-trained optical flow network is used to transfer the knowledge to the NVS student network instead of transferring to the motion vector network. This is a two streams-based learning approach attempting to leverage the advantages of NVS sensing data. However, this work degrades the performance significantly on the experimental datasets due to converting the asynchronous events casting back into synchronising, losing the advantages of neuromorphic sensing data. This work has been improved later in [168] by proposing a graph topology to represent the events. DAVIS240c sensor is used firstly to convert the RGB video sequences into neuromorphic sequences by recording the neuromorphic sensing data from the monitor. Although, this presented framework learning from the event providing a spatial and temporal feature learning, frame-based representation is also generated during the learning to feed them temporal representation into CNN. This frame-based stacking again leads to include redundant information diluting the advantage of reducing the redundancy in neuromorphic sensing data. Table 2.4 summarises the characteristics of all

Table 2.4: Summarizing the main characteristics of event-based neuromorphic-based multi class action recognition systems.

Feature	[146]	[143]	[144]	[168]
Recording scenario	Monitor display	Monitor display	RGB reading	Monitor display
Events generating	Neuromorphic camera	Neuromorphic camera	Emulator	Neuromorphic camera
Camera model	DVS	DVS128	-	DAVIS240c
Learning	Deep learning	Hand-crafted	Deep learning	Deep learning
Classifier	CNN	SVM	CNN	CNN

the existing work on simple behaviour monitoring.

2.7 Datasets

In order to confirm the outperforming of the proposed descriptor, eight challenge action datasets have been chosen. These datasets are commonly and widely used in computer vision research. These datasets are KTH, Weizmann, UCF sports, UIUC1, UCF11, UCF50, HMDB51, and DHA. We also includes the N-Actions dataset which is a native neuromorphic dataset. Table 2.5 shows static information about the datasets that have been used in our experiments and reports the results of accuracy. These datasets have been chosen since it is difficult to find publicly available AAL datasets. At the same time, the human activity is a set of successive actions and in this context we focus on the actions which represents the atoms of the activities. These are the reasons for using these datasets in our experiments. The details of these datasets will be explained in the following.

Table 2.5: Datasets used in our experiments sorted by year of creation.

Dataset	Actions/Activities	Year	#Videos	#Classes	Used in papers	Accuracy %
KTH [4]	Actions	2004	2391	6	[169, 170, 128, 171, 172, 126]	96.8 [126]
Weizmann [114]	Actions	2005	81	9	[140, 173, 174, 175, 170, 176, 177, 178, 179]	100 [179]
UCF Sports [180]	Actions	2008	150	10	[181, 182, 183, 184, 185, 186, 187]	96.22 [187]
UIUC1 [119]	Activities	2008	532	14	[188, 189, 190, 191]	98.9 [191]
UCF11 [192]	Actions	2009	1600	11	[193, 194, 195, 196, 197, 198, 199]	96.94 [199]
UCF50 [200]	Actions	2010	6618	50	[201, 136, 202, 138, 137, 203, 204]	96.4 [204]
HMDB51 [205]	Actions	2011	6849	51	[206, 207, 208, 209, 210, 211]	82.48 [211]
DHA [212]	Actions	2012	532	23	[212, 213, 214, 215, 216, 217]	96.69 [216]
N-Actions [6]	Actions	2019	450	10	–	–

KTH dataset

KTH was proposed by Schuldt *et al.*[4] and showing 6 actions; *i.e.*, boxing, handwaving, handclapping, jogging, running and walking. The dataset is recorded with different cameras and viewpoints. There are 25 different subjects performing the aforementioned actions in four different scenarios. The sequences were captured over homogeneous backgrounds with a static camera recording 25 fps frame rate. Each sequence has a resolution of 160×120 with an average of 4 seconds length.

Weizmann dataset

This dataset was first proposed by Blank *et al.*in [114] and contains 93 video sequences. These sequences have low-resolution sequences of 144×180 with a frame rate of 50 frames per second (fps). The dataset shows nine actors, each of them performing 10 actions, i.e. bend, run, walk, skip, jack, jump, pjump, side, one hand wave and two hands wave. This dataset is widely used in the applications of action recognition.

University of Central Florida (UCF) sports dataset

UCF Sports [180, 183] is a set of action sequences collected from the broadcast channels from various sports containing 150 video sequences with the resolution of 720×480 and and 10 fps frame rate over 10 classes. The actions feature a wide range of scenes and viewpoints. This dataset has been used for the application of computer vision; such as action recognition and action localization.

UIUC1

This is an indoor dataset [119] includes 532 video sequences showing 14 human actions, i.e., walking, running, jumping, waving, jumping jacks, clapping, jump from situp, raise one hand, stretching out, turning, sitting to standing, crawling, pushing up and standing to sitting, captured by a static camera. These 14 actions are performed by 8 actors where each actor does the same action several times. The sequences came with a resolution of 1024×768 and 15 fps frame rate.

University of Central Florida (UCF) 11

UCF11 contains 11 action categories: basketball shooting, biking/cycling, diving, golf swinging, horse back riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog, collected from YouTube. There are around 1600 video clips grouped into 25 groups with more than 4 action clips in it. The video clips have a frame rate of 29.97 fps.

University of Central Florida (UCF) 50

UCF50 is an action dataset including 6676 video clips collected from YouTube. This dataset is an extension of UCF11 dataset to represent 50 real-world human actions, such as, Baseball Pitch, Basketball Shooting, Bench Press, Biking, Billiards Shot,Breaststroke, Clean and Jerk, Diving, Drumming, Fencing, Golf Swing, Playing Guitar, High Jump, Horse Race, Horse Riding, Hula Hoop, Javelin Throw.

HMDB51

The human motion database (HMDB) [205] is one of the largest dataset used to recognise the human motion containing 6849 clips distributed in 51 action classes; each video has $\sim$ 20–1000 frames. The actions categories of this dataset can be grouped in five types based on the body movements. This dataset is considered a challenge due to it includes sequences collected from the Internet and YouTube; therefore, this dataset is a real-world video sequences collection.

Depth-included Human Action (DHA)

DHA dataset was suggested by Lin *et al.*[212]. The dataset consists of 532 sequences compromising 23 action categories performed by 21 subjects (12 males and 9 females). It is recorded using a static Kinect camera in three different scenes with 480×640 resolution. The RGB versions of sequences are used in the experiments.

Neuromorphic Actions dataset (N-Actions)

N-Actions is a dataset captured with a DAVIS346redColor neuromorphic camera. This dataset is firstly presented in [6]. It contains 450 sequences with a resolution of 346×260 to represent 10 human actions using 15 subjects.

2.8 Concluding Remarks

In this chapter, we have outlined the existing work on privacy anonymisation and the relevant contribution on exploring the anonymity domain. In this context, we presented the vision-based sensors that are already used to protect privacy and the relative challenges in each type of sensor. We also explained the exploited solutions to address privacy and exploiting beyond the anonymisation. Furthermore, this chapter introduced a new sensor technology, *i.e.*, neuromorphic camera, which has come up to the application. Based on the existing work, we draw several conclusions:

1. The image filtering methods are considered the standard solution to the preserve the privacy in vision-based monitoring systems. However, the output of these methods is restricted to address privacy and cannot go beyond the privacy for more semantic problems such as action recognition. To overcome this limitation, in Chapter 3 and Chapter 4, we present an anonymity domain that preserves the privacy and provides a useful abstract as a temporal salience data for HAR. The proposed method maintains the action data in each frame as a temporal salience-based silhouette and removes the redundant data. Since these silhouettes are constructed differently based on the actions, discrimination is also formed in these silhouettes. Therefore, extracting the features from these silhouettes improves the accuracy rates for HAR.

2. The video camera is the common vision-sensor that is used widely in monitoring and surveillance applications. This sensor produces frames of intensities-based data including redundancy since the video camera records everything in the field of view (FOV). Processing such intensities consumes the resources of the computer as well as affecting the performance of application. The new technology of vision-sensor, *i.e.*, neuromorphic,

addresses the limitation of recording the intensities by acquiring the change in the intensity instead of the intensity magnitude. This approach reduces the size of the output because the output is in the form of a stream of events. This sensor also addresses the issue of privacy since the output is the orientation of the change in the intensities without any indicating for the intensities.

3. We demonstrate that the standard methods to analysis of the intensity-based data are enabled to process the output of the neuromorphic sensor. Thus, a new method able to explore the neuromorphic domain for more semantic problems, *e.g.*, HAR, is required. Therefore, in Chapter 5, we introduce a new method to extract a meaningful abstract from the events to represent the actions.

Chapter 3

Temporal salience modelling for video-based anonymisation

This chapter presents the proposed method that has been used to increase the level of privacy protection and balancing between the utility and the privacy. It also includes the results of both the subjective and objective evaluations to judge between the proposed method and the existing privacy preservation methods.

3.1 Introduction

This chapter proposes a novel method for video-based identity and appearance preservation and providing an informative anonymity domain. Different hardware and software products have been offered to provide tools for AAL applications. In this context, fusing video-based sensors and computer vision has gained the attention for monitoring human daily living activities and personal wellbeing in AAL applications [14, 218, 16, 219, 220, 221]. Although, these systems and tools have performed well in monitoring, exploring vision sensors for in-home monitoring has often found concerns in protecting privacy [14, 11, 49, 222, 223].

There have been solutions to deal with privacy concerns of video cameras by processing the pixel intensity values spatially to cover the identity details, such as the face or the whole body, by means of masking [62, 63], blurring [224], pixelation [60], etc.. However, after visually anonymising, the utility of such sequences in high level processing, such as, action recognition,

is affected since these methods focus only on the spatial content of the visual data and omit the temporal context information. Such anonymisation methods also cause distortion of the visual data instead of maintaining the visual content. However, these methods require consideration of the trade-off between the visual anonymity and the utility of the anonymised sequences for monitoring tasks [78]. Achieving this trade-off is one of the major challenges associated with using the video camera in AAL. Therefore, maintaining the quality of the obfuscated content is required to improve the reliability of the anonymity domain.

Recently, video-based saliency detection has been proposed to highlight the most dynamic salience content in the video sequences [3, 225, 226, 2, 227]. The outcome of video saliency is a useful abstract for the most dominant visual information in the scene without showing the details since the salient content is represented through highlighting the essential content, simulating perception in the HVS. Visual saliency can be due to the spatial attentive cues as in images as well as due to the temporal saliency due to the motion in a video sequence. Although, salience estimation for video has become a widely addressed topic recently, all methods consider joint spatial and temporal salience modelling. However, since our focus is in the utility, such as HAR, in this chapter we propose a novel temporal salience estimation and demonstrate the use of such salience maps for visual anonymisation. The temporal saliency also seems to be a useful tool for addressing the challenges, such as background clutter often seen in computer vision, since the spatial content is excluded in modelling the temporal salience.

In the case of privacy concealment, the existing filtering-based models lose the accuracy of modelling the most dominant human body parts that are responsible for representing the action due to including the redundant spatial content. Thus, the discrimination among the actions tends to be inaccurate from the perspective of both the HVS and the machine. Therefore, exploring the spatial content to obfuscate the identity leads to inaccurate modelling and misses the discrimination among the actions. Accordingly, utilising the anonymised information obtained by these saliency models seems to be unreliable. This relation between privacy and utility is verified in this chapter from the perspective of HVS, and, in the next chapter, we will test it again from the perspective of the machine.

To address the aforementioned problems, we propose a new temporal salience-based method for video-based privacy preservation. Our proposal is to replace the video sequences with the

computed temporal salience map sequences and then explore the salience sequence for utility tasks, such as, HAR. The computed temporal salience sequences not only capture the temporal events, as in emerging neuromorphic (event-based) cameras [228], but also record significance of those events by means of recording the magnitude of pixel-wise salience in a 0-255 range. Early results of our work were presented as conference papers in [222, 229].

The proposed method captures the motion of the human action in the scene to replace the corresponding spatial information leading to accurate discrimination and a convincing anonymising. This mimics the functionality of emerging neuromorphic (event-based) cameras [228] to capture events by modelling the temporal intensity changes. The anonymised video maps derived from the temporal saliency modelling are further analysed by extracting HOG features for activity recognition tasks. The proposed method provides useful anonymous information which can be further explored in activities of daily living monitoring applications such as action recognition efficiently without making an extra processing, like motion estimation. The main contributions of this work include:

1. Proposing a new temporal saliency based method to increase the accuracy of modelling the saliency in the video sequences and decreasing the complexity of modelling.

2. Proposing a new method to achieve a high level of privacy; and

3. Exploring the anonymised video sequences for highly accurate action recognition [222], *i.e.*, utility.

The rest of this chapter is organised as follows: Section 3.2 reviews the current work on privacy preservation. Section 3.3 presents the proposed method for anonymising video sequences. The performance evaluation of the proposed system for anonymising efficiency as well as utility achievement is presented in Section 3.4 followed by concluding remarks in Section 3.5.

3.2 Related work

Besides the work in this chapter, other anonymity methods have emerged and been presented, which are valuable efforts to preserve privacy. However, contrary to our contribution, these methods are mostly focused on covering the identity silhouette using image processing in the

spatial domain [63, 50, 51, 224] or installing low-resolution sensors [80, 27, 81, 46, 230], where less information for visual recognition is present. Using low-resolution sensors adopts a network of extremely low-resolution cameras [27, 80] or low-resolution colour sensors [81, 46] to capture low-resolution visual images. These sensors have been successfully exploited in the applications of activity recognition [80], behaviour understanding [27] and object localisation [81, 46]. However, these sensors are more sensitive to the local changes in the light conditions [27, 81], which affects the reliability of exploiting the outputs of them in HAR.

The second category of solutions is to adopt the image processing techniques, such as, blocking [50], cartooning [51], Gaussian blurring [224], pixelation [60] and masking with silhouettes [62, 63], to obfuscate the sensitive information. These image filtering-based methods destroy the original intensity magnitudes and include redundant data in the anonymised map. Therefore, exploring the anonymity domains of these methods for HAR affects the accuracy rate of recognising. Furthermore, in these methods, consideration of the trade-off between the privacy protection and utility of the anonymised sequences for monitoring tasks is required [78]. Often, a higher level of privacy protection means a low level of utility and vice versa. This trade-off is one of the major challenges associated with using video-based vision sensors in the application of AAL. Therefore, our proposed approach is a valuable contribution to the development of algorithms to preserve privacy while enabling the subsequent analysis utility tasks, such as HAR.

The current work is a valuable contribution to the development of algorithms to preserve privacy. Nevertheless, only covering the identity details with including a redundancy reduces the usability of the anonymised data for more semantic task, *i.e.*, action recognition. Therefore, modelling the privacy based on the temporal change of the action instead of the intensities seems to provide an informative abstract about the action and presents a useful anonymity domain that can be explored for HAR. Thus, the contribution of this chapter is to provide an anonymity domain considering the obstacles above to address the privacy. The proposed method in this chapter focuses on modelling the action temporally to preserve privacy and presents useful anonymised data for action recognition.

3.3 The proposed method

This section presents the proposed method for estimating temporal visual saliency for visually anonymising the video sequences as detailed in Section 3.3.1 to protect privacy and explore the obtained obfuscated information for action recognition. Figure 3.1 depicts the proposed method to preserve privacy and provide a useful anonymity domain.

3.3.1 Temporal visual salience modelling for visual anonymisation

The proposed method generates the anonymised map of the human action to protect privacy by modelling the temporal saliency in the video sequence. This method depends on distributing the temporal saliency magnitudes based on the most dynamic parts to conceal the privacy, which is crucially attributed in action signature and then action representation. The privacy preservation by modelling the action improves the ability to utilise the anonymised data beyond privacy without using extra information or additional algorithms. The details of modelling the anonymised silhouette of the RGB-based human action are shown in Figure 3.2.

Generating the anonymised map of privacy consists of the following steps. Let s be a video sequence with F frames. First, for every two successive frames, f_t and $f_{t-1} \in s$, where t is the frame index, the frame difference, D_t, is computed to define the change in the pixel intensity over time as

$$D_t(x, y) = f_t(x, y) - f_{t-1}(x, y), \tag{3.1}$$

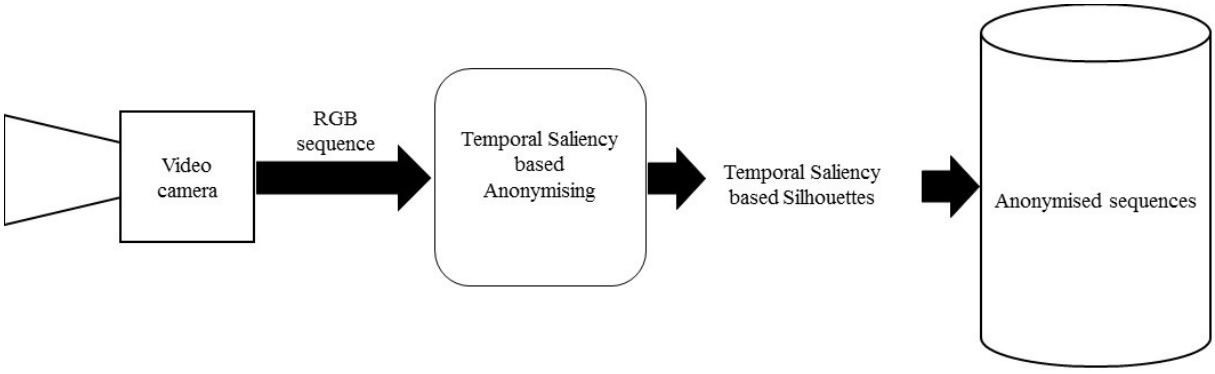


Figure 3.1: Proposed privacy protection model.

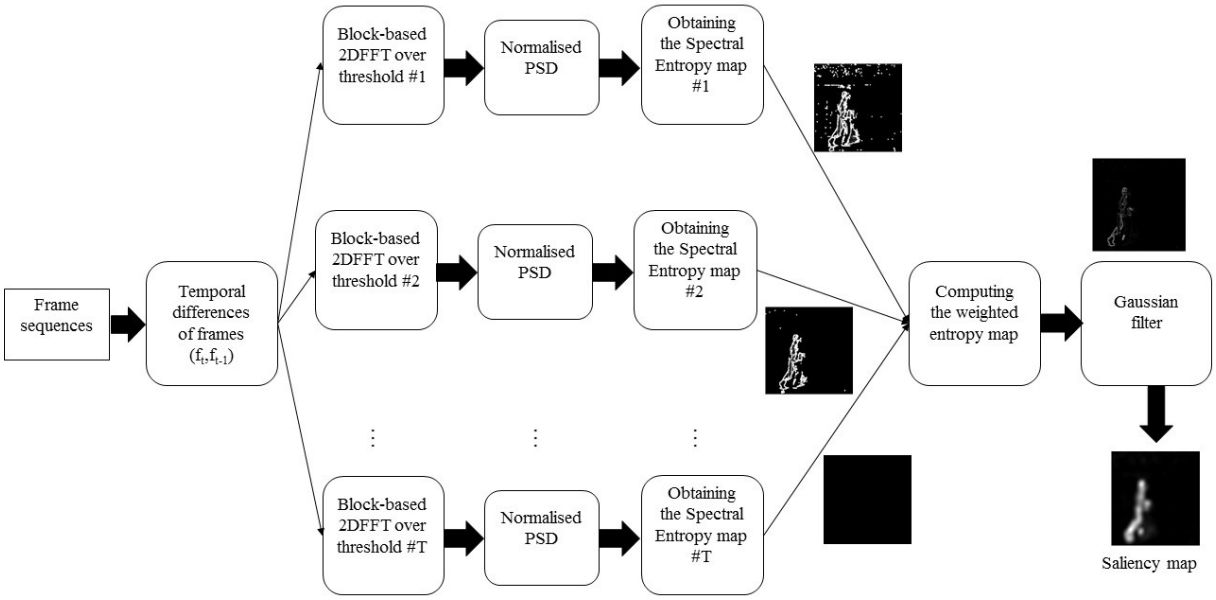


Figure 3.2: Proposed silhouette modelling method based on multiple temporal saliency estimation.

for all (x, y) spatial coordinates in order to measure the magnitude of the movement. The difference at a given pixel can occur for several reasons, for example, illumination change and global motion. Therefore, the frame difference is compared with a user-defined threshold, τ, in order to eliminate the small changes and maintain the dominated moving pixels as follows:

$$\mathbb{D}_t(x, y) = \begin{cases} D_t(x,y) & \text{if } D_t(x,y) \geq \tau \\ 0 & Otherwise \end{cases}, \tag{3.2}$$

where $D_t(x, y)$ and $\mathbb{D}_t(x, y)$ are the frame difference at location (x, y) before and after thresholding with τ, respectively.

Second, $\mathbb{D}_t$ is partitioned into M overlapped blocks, $B = \{b_1, b_2, \cdots, b_M\}$ where each $b_m \in B$ has the size $\beta \times \beta$ and β is odd. Then, for a given block b_m centred at (x, y), a block-based two dimensional fast Fourier transform (2DFFT) is applied on to get the magnitudes of the frequencies of b_m. In order to make up the blocks for pixel at the frame borders, the frame borders are padded with relevant number of zero values according to the chosen β.

Third, the power spectral density (PSD), $\mathbb{S}_{b_m}$, for each block is defined as

$$\mathbb{S}_{b_m}(u, v) = \frac{1}{\beta^2}\mathcal{A}_{b_m}(u, v)^2, \tag{3.3}$$

where $\mathcal{A}_{b_m}(u, v)$ is the magnitude of the 2DFFT coefficient at frequency location (u, v) in block b_m and β^2 is the size of b_m. The PSD is useful to define the distribution of the power of the frequencies in a signal [231]. PSD in our proposed method contributes to describe the distributions of the intensity changes in a region. $\mathbb{S}_{b_m}$ is normalised to suppress the high variation among those in different blocks. This is achieved by normalising with respect to the sum of all PSD components of a given block, such as

$$\hat{P}(u, v) = \frac{\mathbb{S}_{b_m}(u, v)}{\sum_{u=1}^{u=\beta} \sum_{v=1}^{v=\beta} \mathbb{S}_{b_m}(u, v)}, \tag{3.4}$$

where $\hat{P}(u, v)$ is the normalized PSD of $\mathbb{S}_{b_m}(u, v)$. This is followed by the computation of the spectral entropy, $\tilde{\mathcal{E}}$, such as

$$\tilde{\mathcal{E}}_{b_m}(x, y) = \sum_{k=x-1}^{k=x+1} \sum_{l=y-1}^{l=y+1} \hat{P}(k, l) \log(\hat{P}(k, l)), \tag{3.5}$$

where $\tilde{\mathcal{E}}_{b_m}(x, y)$ is the obtained entropy of the element located at the centre of the block b_m and $\hat{P}(\cdot, \cdot)$ is the normalised PSD at $(\cdot, \cdot)$ in b_m. The computation of $\tilde{\mathcal{E}}_{b_m}(x, y)$ captures the contribution of the $\mathbb{D}_t$ values in the neighbourhood of $\mathbb{D}_t(x, y)$. The entropy $\tilde{\mathcal{E}}_{b_m}(x, y)$ is proportional to the amount of variation of magnitudes of the corresponding $\mathbb{S}_{b_m}$. For example, the higher the variation in magnitudes in $\mathbb{S}_{b_m}$ the higher the value of $\tilde{\mathcal{E}}_{b_m}(x, y)$.

This local spectral entropy, $\tilde{\mathcal{E}}_{b_m}(x, y)$, fairly captures the variations in $\mathbb{D}_t$ to identify the temporal salience in a frame. It exploits the source of the most dominant intensity changes to model the underlying motion (with respect to the action). Most of the time, it is difficult to determine the perfect value of τ in Eq. (3.2) to maintain the desired changes and suppress other noisy changes because the motion levels vary according to the actions in sequences. To make this representation more robust and generalised, we further vary τ by defining a set of thresholds, $\tau_h = 2^h$, where $h = 1, \cdots, \mathcal{N}$, with maximum number of user defined threshold levels, $\mathcal{N}$. For each pixel location (x, y), a set of entropy values, $\tilde{\mathcal{E}}^{\tau_h}_{b_m}(x, y)$ for the corresponding block, b_m, considering all τ_h is computed as

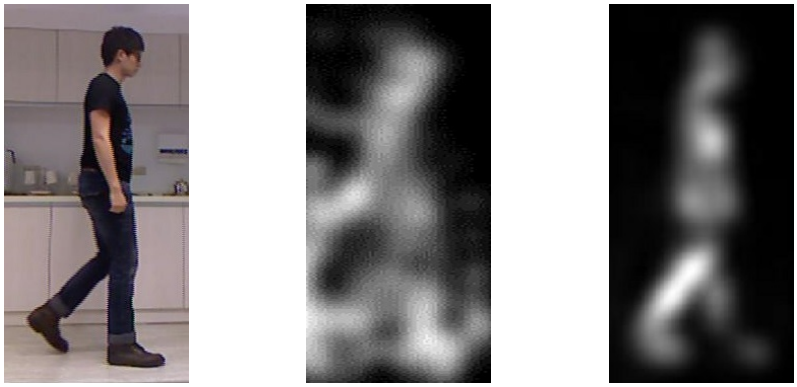
$$\tilde{\mathcal{E}}^{\tau_h}_{b_m}(x, y) = \sum_{k=x-1}^{k=x+1} \sum_{l=y-1}^{l=y+1} \hat{P}^{\tau_h}_{b_m}(k, l) \log(\hat{P}^{\tau_h}_{b_m}(k, l)), \tag{3.6}$$

where $\hat{P}_{b_m}^{\tau_h}$ are the normalised PSD of block b_m, respectively, according to the threshold τ_h. Finally, the weighted entropy, $\hat{\mathcal{E}}(x,y)$, across all entropy maps, $\tilde{\mathcal{E}}_{b_m}^{\tau_h}(x,y)$, over all $\mathcal{N}$ thresholds is computed as

$$\hat{\mathcal{E}}(x,y) = \frac{\sum_{h=1}^{\mathcal{N}} \tau_h \tilde{\mathcal{E}}_{b_m}^{\tau_h}(x,y)}{\sum_{h=1}^{\mathcal{N}} \tau_h}. \tag{3.7}$$

$\hat{\mathcal{E}}$ highlights the silhouette of the foreground object by modelling the distribution of the saliency based on the relation between the action and the human body parts. Furthermore, this approach is considered a crucial factor in the distinction between actions since the magnitudes of the entropies will be distributed across the silhouette based on the action class. The $\hat{\mathcal{E}}$ map is normalised to be in the range of grey level values in the range [0 255] and smoothed by applying a 2-D Gaussian kernel in order to fill in the small holes and obtain the final spectral entropy based temporal visual salience map based silhouette, $S_{\hat{\mathcal{E}}}$, for privacy preservation. It links the neighbouring pixels that are close to each other to construct the temporal silhouette region.

Figure 3.3 shows an example of generated silhouettes using the proposed method. It demonstrates the benefit of using multiple thresholds to compute the weighted entropy, $\hat{\mathcal{E}}(x,y)$. It can be seen in Figure 3.3 (c) that the generated silhouette further highlights the most dynamic body parts used in the action compared to the rest since the moving parts are represented with



(a) (b) (c)

Figure 3.3: Anonymising modelling of frame $\#15$ from the walking sequence of the participant $\#1$ in DHA dataset: (a) original frame, (b) single threshold based anonymising Eq. (3.6) with a single threshold and (c) Silhouette modelling using Eq. (3.7) with multiple thresholds.

high temporal visual salience magnitude values. These moving parts are represented by increasing the highlighting of their temporal salience magnitude using the proposed approach of anonymity. Otherwise, if we depend only on a single threshold formula, Figure 3.3 (b), affects the accuracy of modelling the silhouette of the action since the human body parts seem to have the same magnitude without focusing on the main moving part. Instead, with the proposed local spectral-entropy based obfuscating method, the variation in movements will be reasonably modelled. The proposed method protects the privacy as well as maintaining the most useful information about the action. More examples have been displayed in Figure 3.4 to show the performance of the proposed method. This modelling can help to discriminate among the actions from using their temporal salience maps without extra information.

In Figure 3.4, we notice that the silhouette for a specific action is changed with time based on modelling the action over time. This modelling depends on the amount of motion that has been generated from each part of the human body at a specific time. For instance, in the case of jacking action, third row, the silhouette has a different pattern every time, as some parts are attenuated, and others gain extra highlighting. In addition, the algorithm generates different saliency maps for one-hand waving and two-hands waving actions, as we can see in *row 4* and *row 6*, respectively, since the patterns of these two actions are different. This representation is important to create a useful abstract that can be useful to extract the action description by accurately identifying the variation between the actions.

The examples in Figure 3.4 show that the proposed method can be used for two useful purposes: protecting the privacy efficiently by obfuscating the essential information in the scene and eliminating unnecessarily redundant information. The output of the proposed method makes the video-based sensor less intrusive and more acceptable in the real world. Furthermore, modelling the anonymity in the form of saliency maps provides useful abstract to be explored by the descriptor to extract powerful discriminating features and achieve efficiency in the action recognition application.

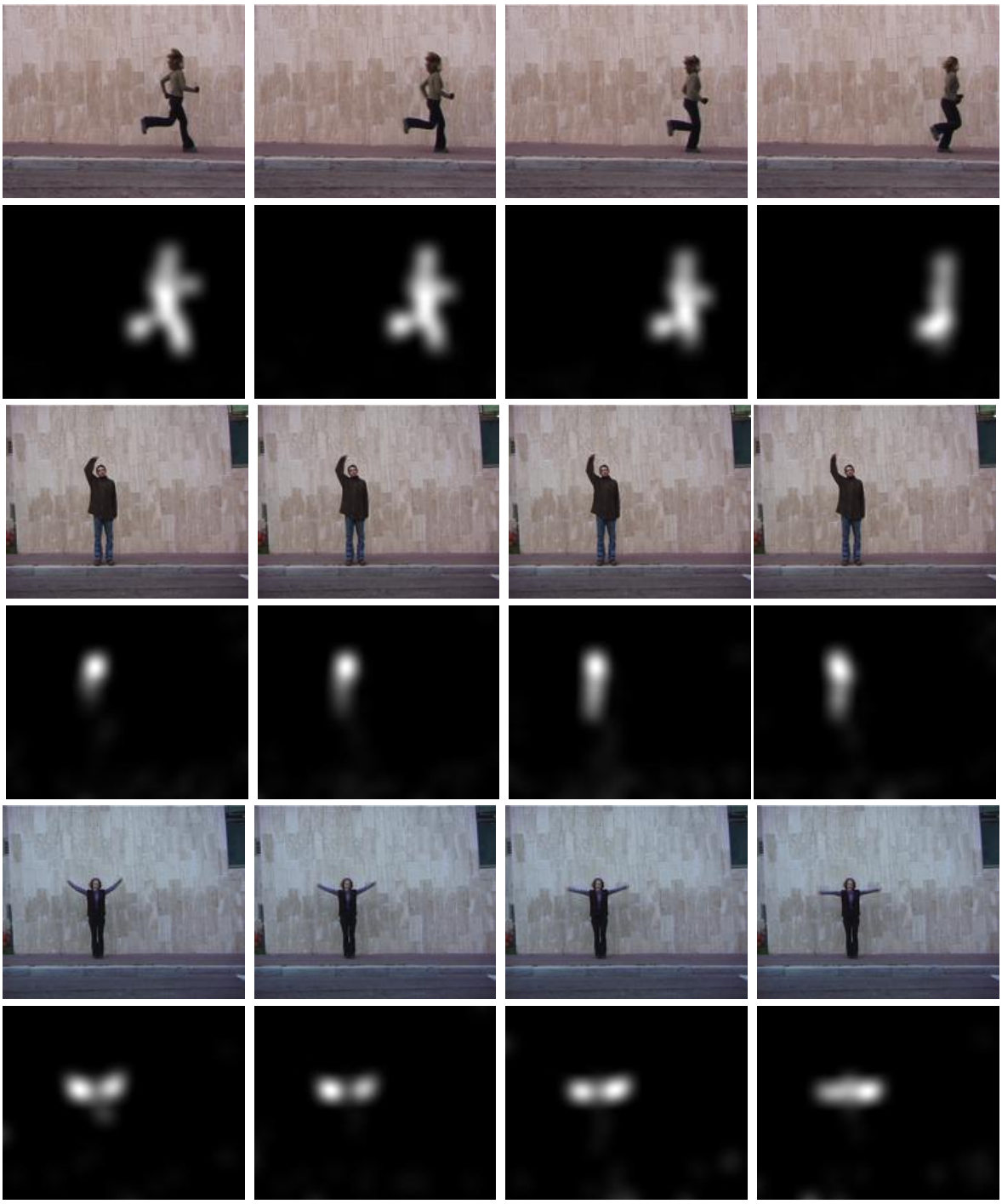


Figure 3.4: Temporal saliency based silhouettes for six actions from Weizmann dataset: *rows 1* & *2* are the original frames and their corresponding temporal salience maps for run, *rows 3* & *4* are the original frames and their corresponding temporal salience maps for one hand waving, and *rows 5* & *6* are the original frames and their corresponding temporal salience maps for two hand waving. Rows 2, 4, and 6 show how the silhouettes are changed over time for these three actions.

3.4 Performance evaluation

For the evaluation of the proposed video-based anonymisation method, we use both subjective and objective approaches to evaluate it. In this evaluation, we explore four publicly available datasets, *i.e.*, DHA [212], KTH [4], Weizmann [118], and UIUC1 [119]. These datasets are common and widely used in computer vision applications. They are also considered challenging datasets due to the variation in the environment (outdoor and indoor), the usage of single and multiple cameras and the variation in participants race and age groups. In the beginning, the proposed temporal salience based method is evaluated by objective metric and comparing it to the existing video-based saliency methods followed by the anonymisation subjective evaluation.

3.4.1 Experiment setup

In our experiments, we use $\beta = 3$ and $h = 7$ for evaluating the proposed visual anonymization algorithm. The weighted entropy maps, $\hat{\mathcal{E}}(x, y)$, are smoothed using a 2D Gaussian kernel with $\sigma = 6$.

3.4.2 Temporal salience evaluation

The performance of the proposed temporal salience-based detection method is evaluated by conducting an objective evaluation using the Area Under Curve (AUC) values of Receiver Operating Characteristics (ROC), which is the most widely used metric when comparing different saliency models. The AUC is used to evaluate the performance of the saliency models by considering saliency map as a binary classifier of fixations at various threshold values [232]. AUC is adopted to compare the performance of the proposed method with video-based saliency state of the art, Fang *et al.*[3], Kim *et al.*[2] and Wang *et al.*[225] in terms of the accurate salient region detection and the time of computation.

The experiments have been conducted on DHA, Weizmann, and UIUC1 datasets. The result of AUC is depicted in Table 3.1 for each method. This table also includes the average AUC and the average execution time for each method at the bottom of this table. These results show that the proposed method, which only models temporal salience, has comparable accuracy in terms of AUC with the existing methods, while taking low computational time. The reason for this

outperforming is that AUC calculating depends on comparing the saliency map with the ground truth, which means that the saliency maps include a redundancy. This redundancy impacts on the process beyond the saliency computing, *e.g.*, feature extraction, because the discrimination among the actions can be reduced (we will prove this argument in the next chapter). Most of the redundancy comes from including the spatial-saliency in computing the final saliency map. Therefore, depending on the temporal salience data reduces the redundancy and improves the saliency-based feature extraction.

Table 3.1 also shows the average execution time per frame which is calculated based on collecting the execution time for all frames across all the datasets in accordance with the specification of the PC that will be explained in Section 3.4.5. Then, the average is taken to represent the execution time for modelling each saliency map. We notice that Wang and the proposed method have less computational complexity compared to Fang and Kim. However, the proposed method produces AUC value better than Wang, which makes it a reliable candidate in real-life applications. These average AUC values and the corresponding execution time are also shown in Figure 3.5 to illustrate the performance of each method in terms of modelling the silhouette and the required time to achieve this modelling.

Examples of salience maps for various action sequences from DHA and Weizmann datasets using our proposed method and existing work are shown in Figure 3.6. These maps represent different human actions that are obtained by using the methods in Table 3.1. In Figure 3.6, we observed that the differences in modelling the actions whenever using our temporal saliency maps and those computed using the existing video-based saliency models. For instance, in *row 3*, hand clapping activity where the furniture of the room has been highlighted more than the hands that act during the action.

Table 3.1: Average AUC and the corresponding execution time of the proposed method and state of the art.

Sequences	***Fang***	***Kim***	***Wang***	***Proposed***
DHA	0.87	0.92	0.82	0.91
Weizmann	0.95	0.96	0.95	0.95
UIUC1	0.98	0.98	0.98	0.92
Average AUC	0.93	0.95	0.92	0.93
Average time (sec)	31.8	34.4	4.21	5.4

Furthermore, in other cases, the existing methods highlight the human body parts that are irresponsible for the action, *e.g.*, *row 4*, and obfuscate the responsible parts; such as running and walking activities. This modelling makes it difficult to obtain meaningful features from these maps and confuses both the machine and the HVS to discriminate among the actions. It is evident that our proposed salience maps only capture the body parts relevant to the action, whereas, other methods capture other spatial information and the full body which are not relevant to the action.

3.4.3 Subjective evaluation of the proposed temporal salience-based anonymity

We evaluated the effectiveness of the proposed visual anonymisation using human observers. A survey with 30 individuals participants was conducted to evaluate the proposed method and state-of-the-art filtering algorithms for visual anonymisation. In this survey, the participants were divided into four groups, where each group evaluated a specific anonymised dataset using the proposed methods and the existing methods. The datasets of DHA, KTH, Weizmann, and

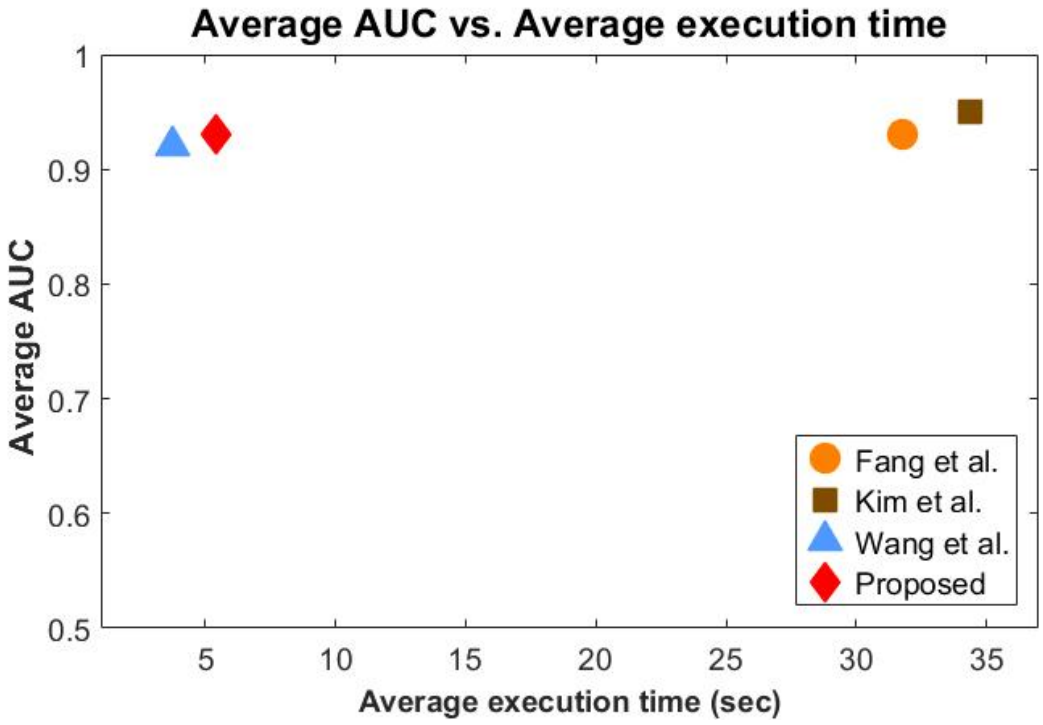


Figure 3.5: Average AUC and the execution time per frame measured by seconds for video-based saliency state of the art and the proposed method.

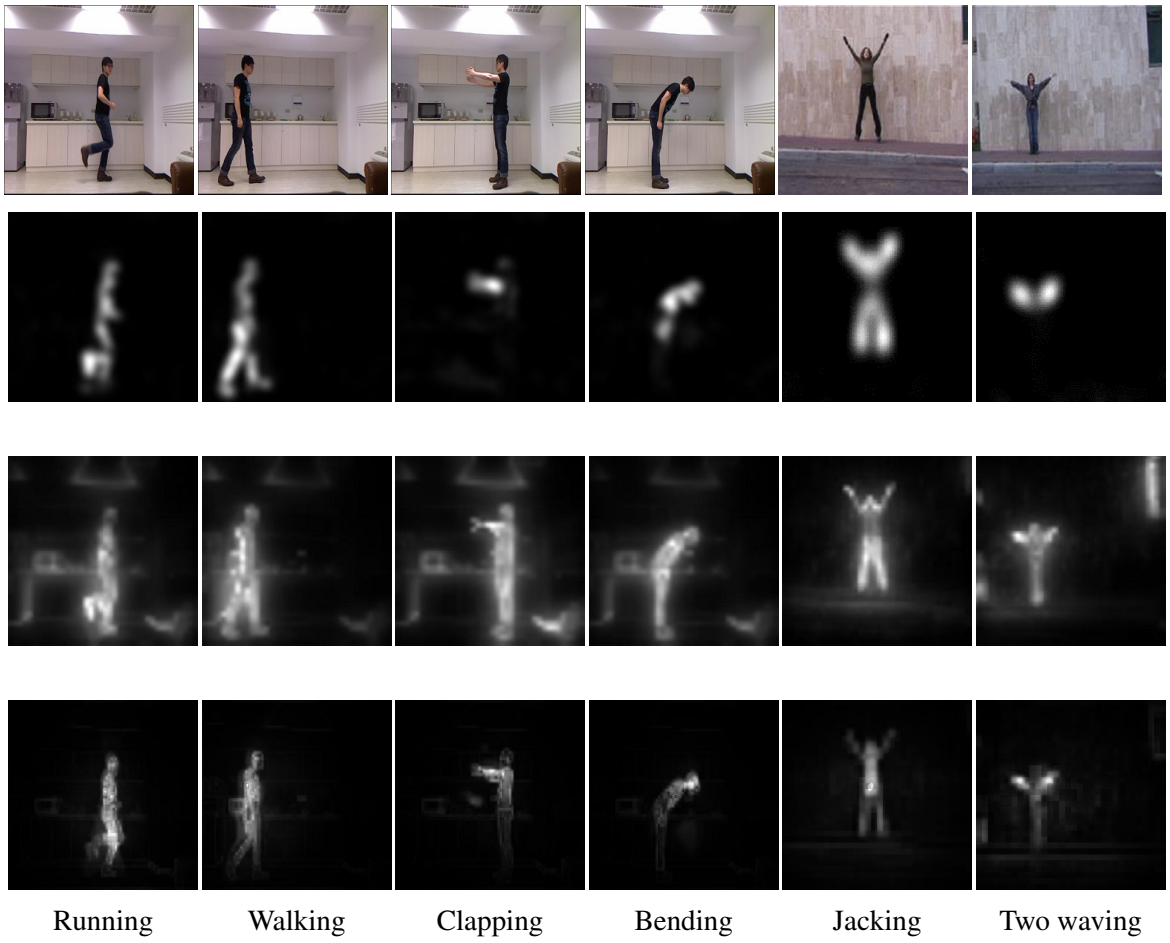


Figure 3.6: Representing the action using the video-based salience detection for a set of human actions. *Row1*: original RGB frames, *Row2* corresponding temporal salience maps of our method, *Row3* corresponding spatio-temporal saliency maps using method in [2] and *Row4* corresponding spatio-temporal saliency maps using method in [3]. The first fourth columns have four actions from DHA dataset while the last two columns have two actions from Weizmann dataset.

UIUC1 were used in this subjective evaluation since they are recorded using a static camera, and our method is proposed to use a static camera as well.

In total, 108 anonymised video sequences for different actions were selected equally from five methods (blurring with $\sigma = 5$, blurring with $\sigma = 8$, pixelation, solid silhouette and binary silhouette) and the proposed method. These sequences have been spread out into four groups and each group was allocated to separate a set of participants for evaluation. Table 3.2 shows information of each group of evaluation and the number of sequences that have been assigned to each group. For each anonymised sequence, the participants were asked to answer two questions

Table 3.2: Evaluation groups and their information.

Dataset	No. of video sequences	No. of participants	Participants' details
DHA	30	8	• gender (7 Males and 1 Female) • age (30-35) • experience (academic)
KTH	24	7	• gender (5 Males and 2 Females) • age(30-55) • experience (academic and non-academic)
Weizmann	24	7	• gender (7 Males) • age (30-35) • experience (academic)
UIUC1	30	8	• gender (6 Males and 2 Females) • age(30-40) • experience (academic and non-academic)

about what they were able to see in the sequence.

Figure 3.7 shows an example of a few selected frames from the sequences that are used in the survey. Concerning the used sequences, we selected a set of anonymised action sequences for different living daily human actions obtained from the datasets mentioned above. In the next sections, we describe the planning for this subjective evaluation and present the obtained results and the explanation of them.

Table 3.3: The questions and their corresponding answers.

#	Question	Possible Answers
1	Which of the following activities better match the video sequence?	[walking, running, jumping, standing, waving, $\cdots$, I don't know]
2	How well the person is anonymised in the video sequence?	[0 (not anonymised), . . ., 5 (perfectly anonymised)]
3	Can you recognise the following features from the video sequence?	tick one or more choices from the following (gender, age group, face, hair, clothes, race, and non).

Planning of the subjective evaluation

The purpose of the survey is two fold. Firstly it aims to find out the effectiveness of the proposed method's visual anonymisation. Secondly, to evaluate whether the utility of the video is affected due to the anonymisation. In this case the utility was considered as the ability for an observer to accurately recognize the action present in the sequence. Three questions, shown in Table 3.3, were included in the survey to achieve these two purposes.

The first question aims to evaluate the level of visual anonymisation achieved by a particular method as perceived by the observer. They are asked to score the level of anonymity on a

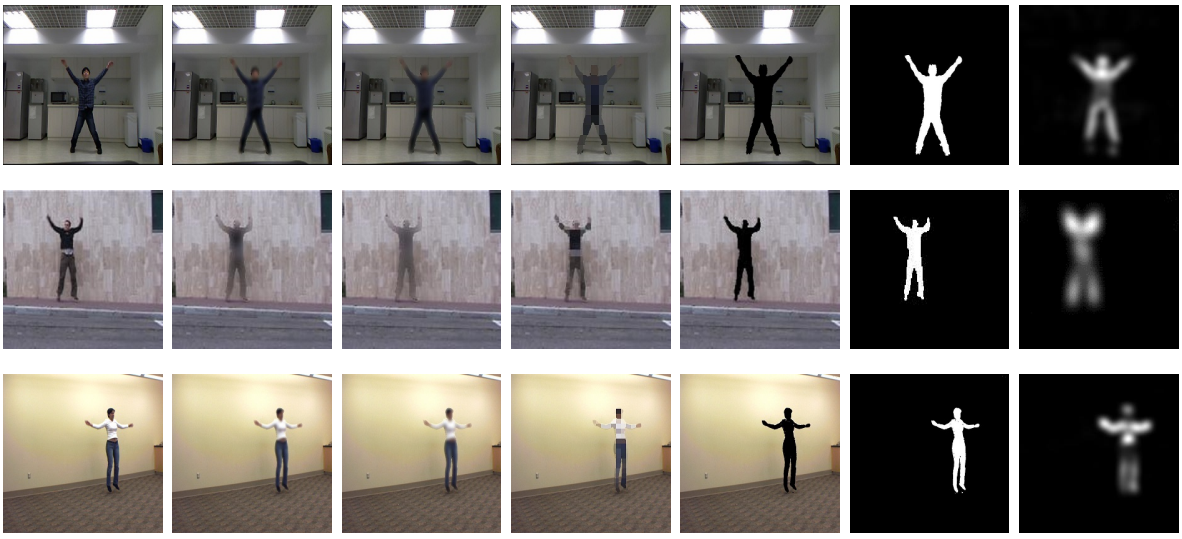


Figure 3.7: Example of the anonymisation state of the art and the proposed method for sampled frames from three datasets: *row1* DHA, *row2* Weizmann, and *row3* UIUC1. The columns from left to right: original, blurring with $\sigma = 5$, blurring with $\sigma = 8$, pixelation, silhouette, binary and the proposed method, respectively.

discrete scale from 0 (no anonymisation) to 5 (perfect anonymisation). The score is regarded to which one they thought could provide enough protection and reduce the concern about privacy protection. The second question collects the identity attributes, such as, gender, apparent age, facial features, clothes, hair and race, that can be recognised by the participants. These attributes are considered sensitive information that has to be protected by a visual privacy preservation model. The unmeasurable attributes were not considered due to the difficulty to determine them in the visual domain. The response to this question needs to be compatible with that for the first question. For example, a score of 5 for the anonymisation level means none of the identity clues can be recognized from the anonymised video. Finally, the third question estimates the ability of anonymisation method to retain useful information that can be used to identify the human action present in the video. This quality relies on the level of anonymity. In other words, if we need to increase the anonymity, the quality of the information has to be discarded and vice versa. The participants were asked to label the action presented in the obfuscated sequence using the information that was retained in the concealment model.

At the beginning of a survey session, the purpose of the evaluation is conveyed to the survey participants. The region of anonymity of a scene is restricted to the human in the scene, but not for the background. The test video set used in the survey consists of various people performing various actions. We aimed to minimize the repetition of the same person doing different actions. Using the same video sequences with versions can help the participants to use their memory to recall the missed details and/or biased to the same answer ignoring the difference between the models. However, in a few cases we use two different models for the same sequence in order to analyse the ability of the participants to recognise between them and if the method can make the difference for the participant or not.

Concerning the appearance clues identification, a list of measurable variables, *i.e.*, a person's visual clues that take specific values in a determined domain, such as the gender is measurable in the domain $[male, female]$, and so on. The unmeasurable clues have been discarded due to the difficulty to determine the domain for these features.

As we mentioned before, the number of the video sequences in this evaluation is 108 sequences distributed as follows: DHA=30, UIUC1=30, KTH=24 and Weizmann=24. The number of video sequences that have been used in the evaluation depends on the size of the dataset

Table 3.4: Number of questions collected from the datasets.

Dataset	Number of answers
DHA	720
KTH	504
Weizmann	504
UIUC1	720

and the number of actions in each dataset. Thus, the number of the video sequence is distributed among the anonymisation model, which is six models except for KTH dataset, where there are four models. The evaluation results in a total of 324 questions (three per sequence) to collect the responses of this evaluation.

In this survey, we avoid using the same sequence that has been tested by the anonymised model in the evaluation as much as possible and make the samples of the video sequences diverse. Using the same video sequences with versions can help the participants to use their memory to recall the missed details and/or biased to the same answer ignoring the difference between the models. However, in a few cases we use two different models for the same sequence in order to analyse the ability of the participants to recognise between them and if the method can make the difference for the participant or not. We also distributed the video sequences into four groups based on the dataset as aforementioned in order to use as much sequences as possible, and, at the same time, this way can avoid the biasing to one dataset. Table 3.4 shows the number of the answers that have been collected for each dataset.

Subjective experiment results analysing

The results of the subjective evaluation have been presented in Figures 3.8-3.21. These figures aim to interpret the results and make it easy to understand. The first part of the evaluation, where the visualisation identity concealment models are used to evaluate the utility the viewpoint of HVS, is shown in Figures 3.8-3.11 for the DHA, KTH, Weizmann and UIUC1 datasets, respectively. The correct answers for each anonymisation model are collected and the average amount, Av, is calculated as

$$Av_d = \frac{A_d}{P_d \times G_d}, \tag{3.8}$$

where A_d, P_d and G_d are the total correct answers, number of participants and the number of ground truth answers for each anonymisation method, respectively, for the dataset d. We used equal number of anonymised sequences and thus there is also equal number of ground truth for all datasets.

Figures 3.8-3.11 show the average values of the correct participants' answers for six anonymisation models, except KTH dataset where four anonymisation methods were evaluated, as Silhouette and Binary masks were not available for the actions in the KTH dataset. The overall average of the results in these figures is shown in Figure 3.12. In this figure, we can observe that the average value of some anonymity methods is better than the proposed method. On the one hand, for instance, blurring model with $\sigma = 5$ in Figure 3.12 seems to achieve better results from the viewpoint of the participants. On the other hand, this means that the quality of the anonymity is low because this method considers the trade-off between privacy and utility. These obtained results explain the dependency of each method in considering this trade-off. However, the judgement of the superiority of any model will be suspended until the results of the anonymity are interpreted.

In general, in Figure 3.12, we observe that the methods of blurring and silhouette seem to

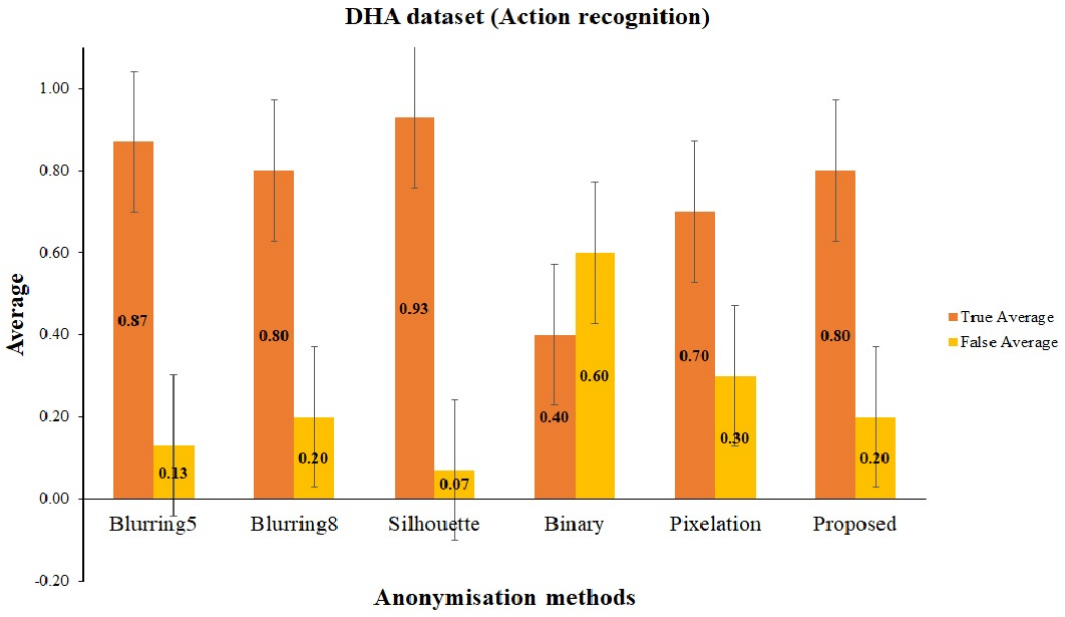


Figure 3.8: The average activity recognition considering all the participants' response for the DHA dataset.

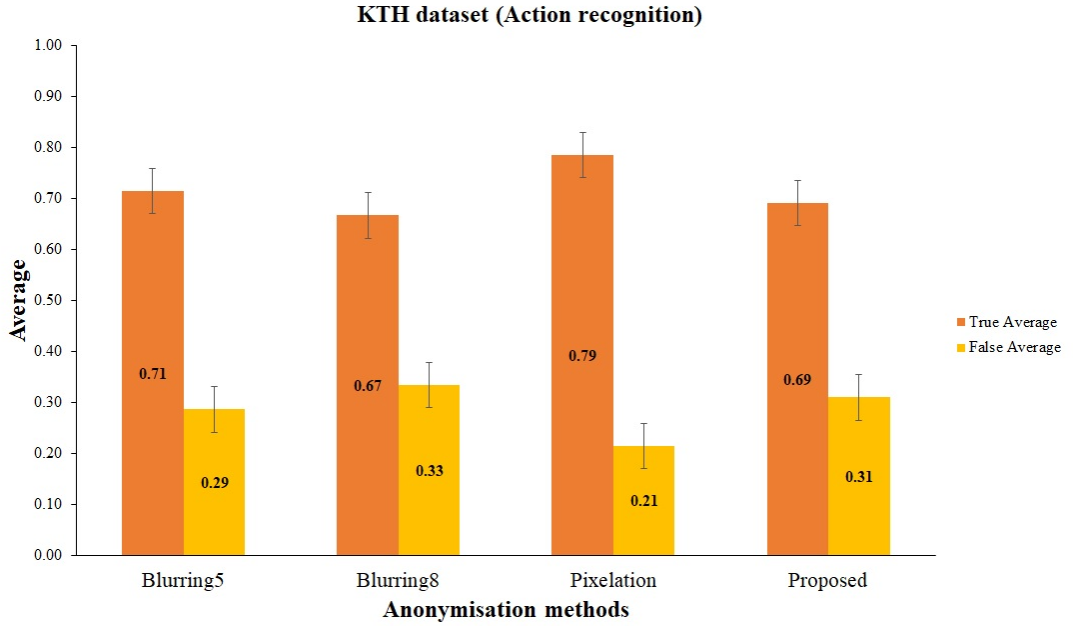


Figure 3.9: The average activity recognition considering all the participants' response for the KTH dataset.

have the best action identifying despite low levels of protection. Furthermore, we can see that the results of these models lose the stability due to the quality of the visual content in each

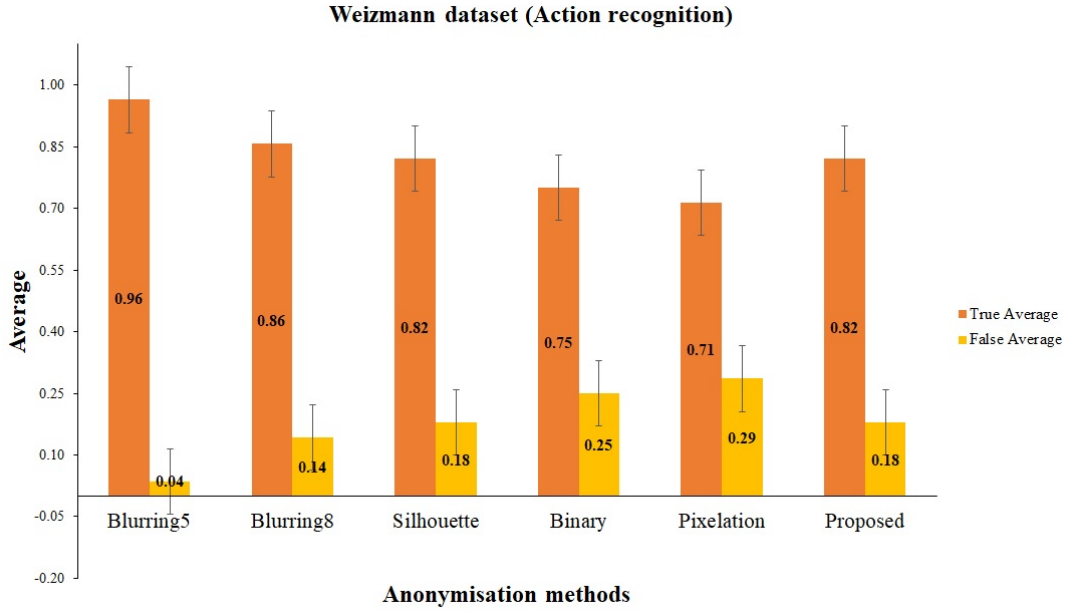


Figure 3.10: The average activity recognition considering all the participants' response for the Weizmann dataset.

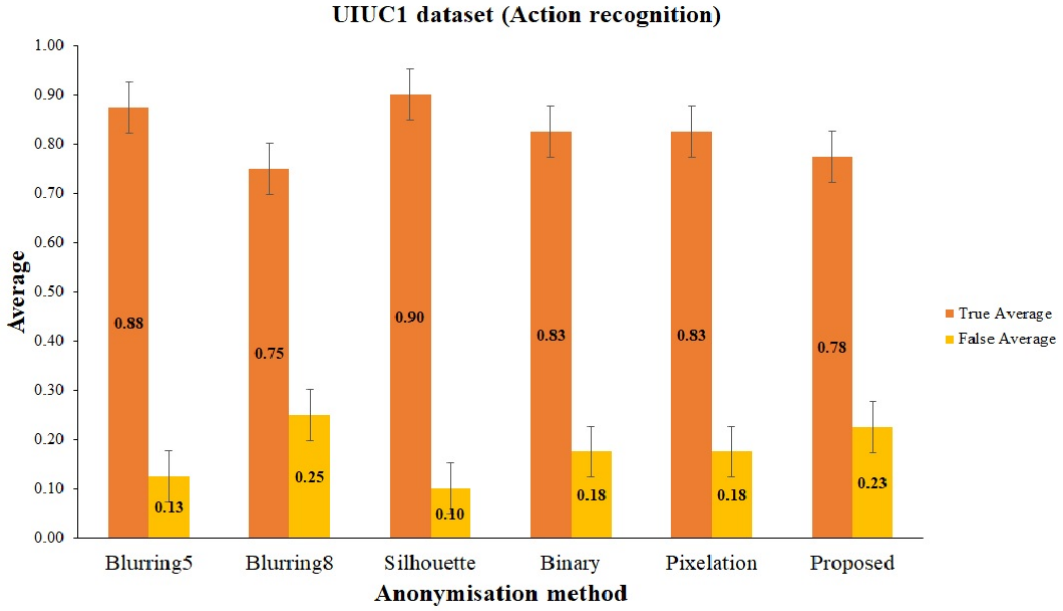


Figure 3.11: The overall average activity recognition considering all the participants' response for the UIUC1 dataset.

dataset. For instance, in KTH dataset, the blurring presented a low level of action recognition compared to the pixelation, while the same method is the best one in the case of using Weizmann dataset. This fluctuation in the results is due to the variation in the quality of visual data of the datasets.

For the protection level that meets the desires of the participants, Figures 3.13-3.16 show the relevance the participants assigned to the privacy protection for DHA, KTH, Weizmann and UIUC1 dataset, respectively, on a scale from zero, non-anonymity, to five, perfect anonymity. The average amount has been interpreted using the following formula:

$$Av_d = \frac{A_d}{P_d \times V_d}, \tag{3.9}$$

where A_d, P_d and V_d are the total correct answers based on the responses of the participants, number of participants and the number of video sequences that are used in the survey, respectively, for the dataset d.

The interpretation of the participants' answers indicates that the proposed method achieves better performance than the filtering algorithms in terms of concealment of the privacy and all relevant attributes. The results in Figures 3.13-3.16 confirm that the proposed method of

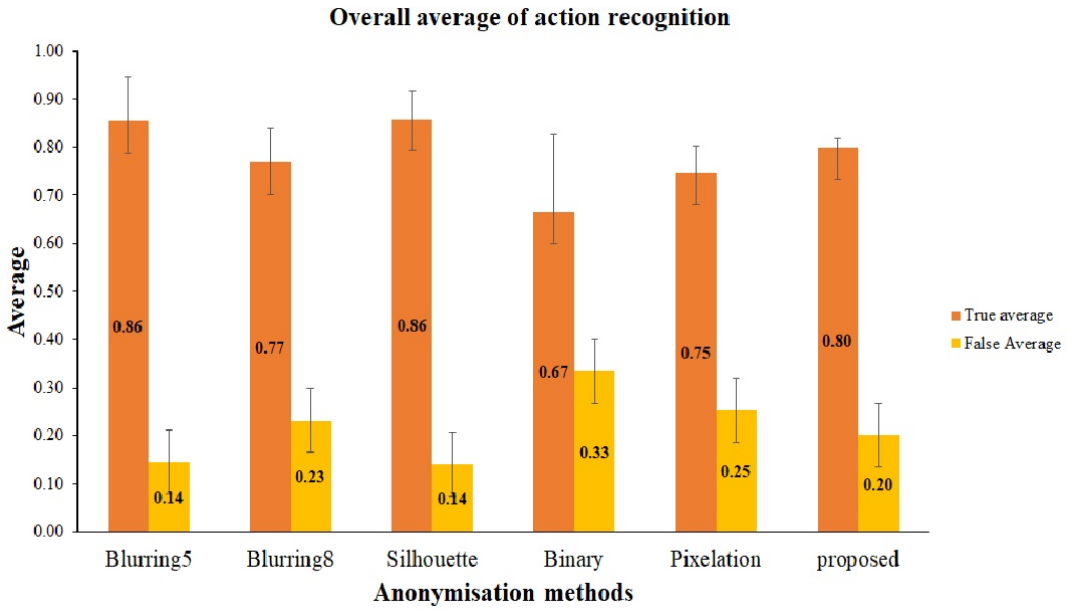


Figure 3.12: The average activity recognition considering all the participants' response for all datasets.

anonymisation achieves the perfect anonymisation, reaching to a high level of privacy protection from the viewpoint of the participants. The reason of this achievement of the proposed method is due to modelling the action using the temporal information instead of covering the silhouette spatially. This reasoning conclusion can be observed from the average scoring in Figure 3.17 of

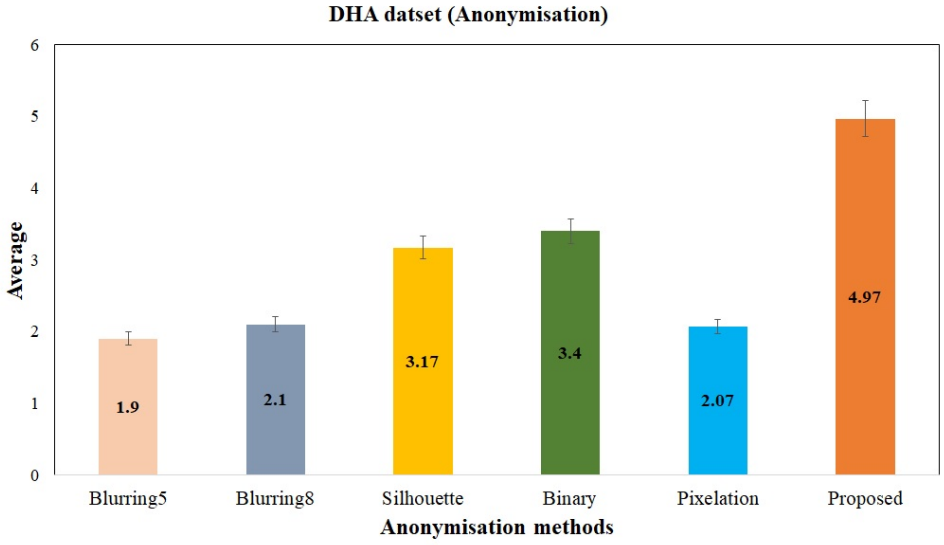


Figure 3.13: The average of anonymisation degree considering all the participants' response for the DHA dataset.

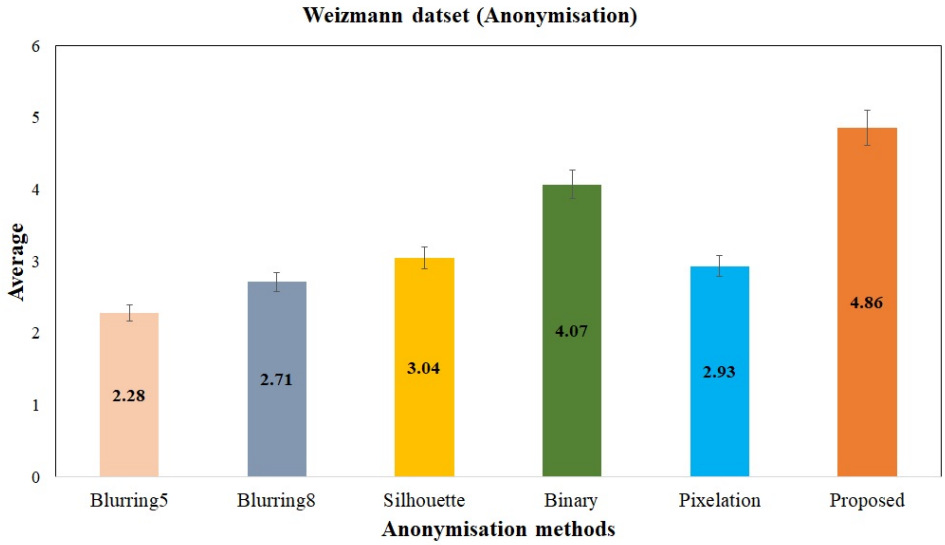


Figure 3.14: The average of anonymisation degree considering all the participants' response for the Weizmann dataset.

the participants' answers for the proposed method compared to the state of the art anonymisation methods. We can also observe that non-intensity-based methods, *i.e.*, binary and the proposed methods, achieve intensity-based algorithms. The low-level of privacy based on blurring and pixelation is because these methods include the intensities in the anonymity domain and these

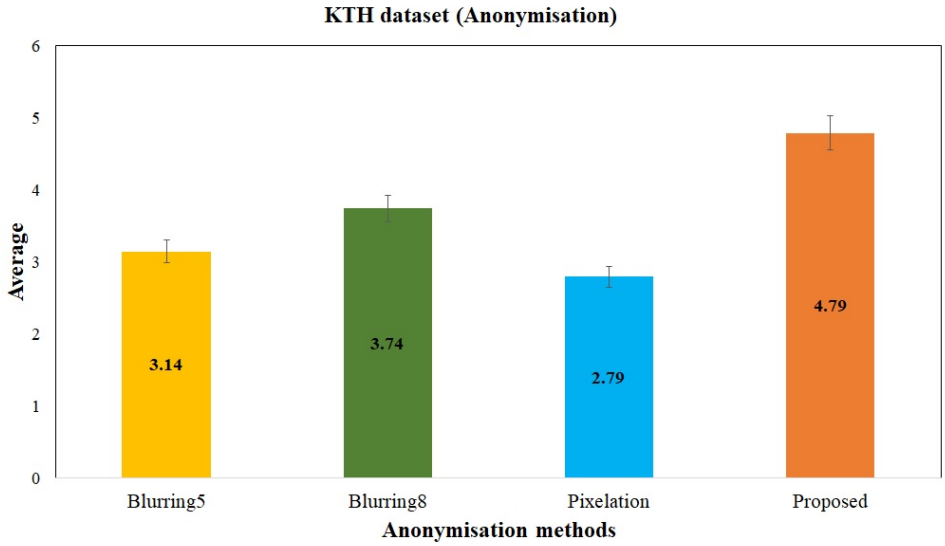


Figure 3.15: The average of anonymisation degree considering all the participants' response for the KTH dataset. The average degrees for Silhouette and Binary methods are excluded because the ground truth maps are unavailable for this dataset that can be used for getting the binary masks and forming the silhouettes.

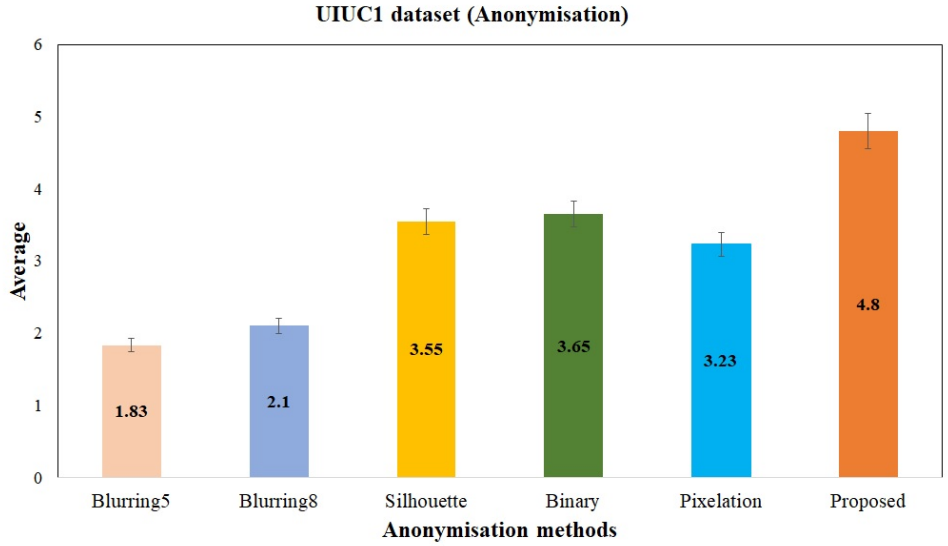


Figure 3.16: The average of anonymisation degree considering all the participants' response for the UIUC1 dataset.

intensities are used to identify privacy.

Finally, the third part of the evaluation focuses on identification of a set of appearance attributes, *i.e.*, gender, age group, clothes, hair, facial and race. Bi-charts in Figures 3.18-3.21

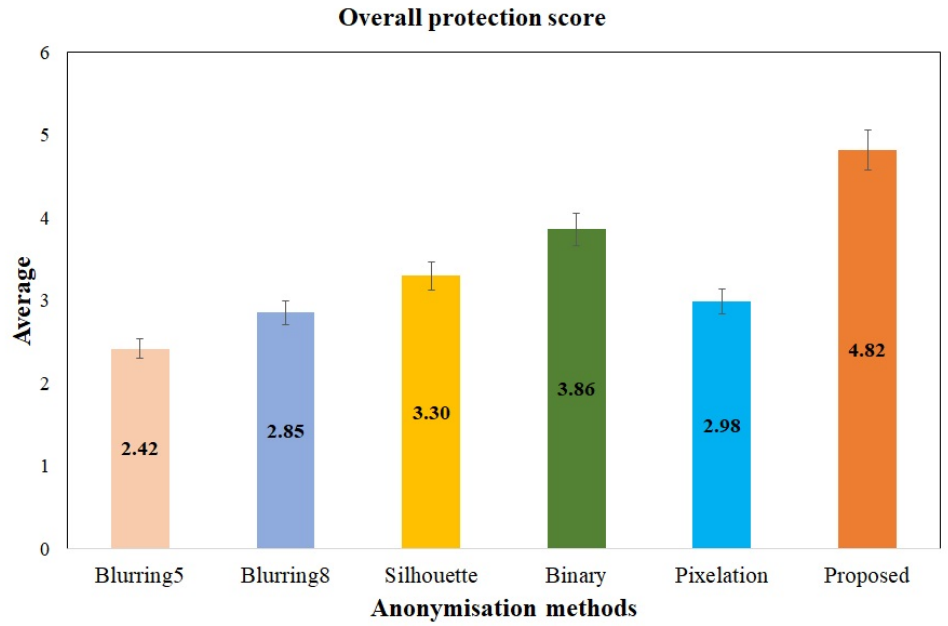


Figure 3.17: The overall average of anonymisation degree considering all the participants' response for all datasets.

illustrate the percentages of recognising the appearance clues in DHA, KTH, Weizmann and UIUC1 datasets, respectively. These clues are considered sensitive information that has to be protected by the vision privacy preservation model. The responses of the participants show that the proposed method of anonymity conceals the attribute of human appearance completely. The perfect anonymisation using the proposed method is a result of depending on the temporal salience modelling leading to model the action across time instead of modelling the body of the human.

A similar result is obtained by using a binary silhouette method. However, our method outperforms the binary silhouette because the binary silhouette method uses the spatial information to construct the silhouette which means that the appearance details, such as the hairstyle, for example, can be used to recognise the gender of the person. In Figures 3.18 (f)-3.21 (f), the percentage of the difficulty to recognise the appearance attributes is the highest compared to the filtering algorithms. Thus, the proposed anonymisation method achieves between $89\% - 100\%$ of concealment, which is the highest compared to the current work. This high level of anonymisation proves that temporal modelling achieves better compared to spatial modelling.

At the end of this subjective evaluation, we conclude that converting the visual data into a useful abstract is more informative, *i.e.*, intelligible, and secure, *i.e.*, anonymised. This conclusion can be illustrated in Figure 3.22, where each method used in the subjective evaluation is located in this figure based on satisfying both privacy and utility scores using the average answers of the participants. It is clear from Figure 3.22 that the proposed temporal salience method achieves the highest level of privacy and outperforms the existing methods. We also notice that blurring with $= 5$ achieves a high level of utility. The reason for blurring is because this method has a low level of obfuscation; therefore, it is easy to recognise the action. If we increase the level of privacy for blurring, the level of utility is reduced. This trade-off can be noticed whenever comparing the results of two blurring methods in Figure 3.22.

Besides, the binary mask-based anonymity gives a high privacy preservation level compared to the first four methods. However, the obtained privacy based on the binary mask is less than the proposed method because the binary mask creates a silhouette with sharp edges that can explain some details of the appearance, such as the gender.

The results in Figure 3.22 indicate that the proposed method obtains a higher level of pri-

Figure 3.18: Percentages of recognising the appearance attributes in DHA dataset using each method: a comparison.

vacy protection and produces informative domain that can be explored for action recognition. Outperforming the proposed method for achieving both privacy and utility at the same time encourages the researchers towards leverage of using the vision camera for AAL.

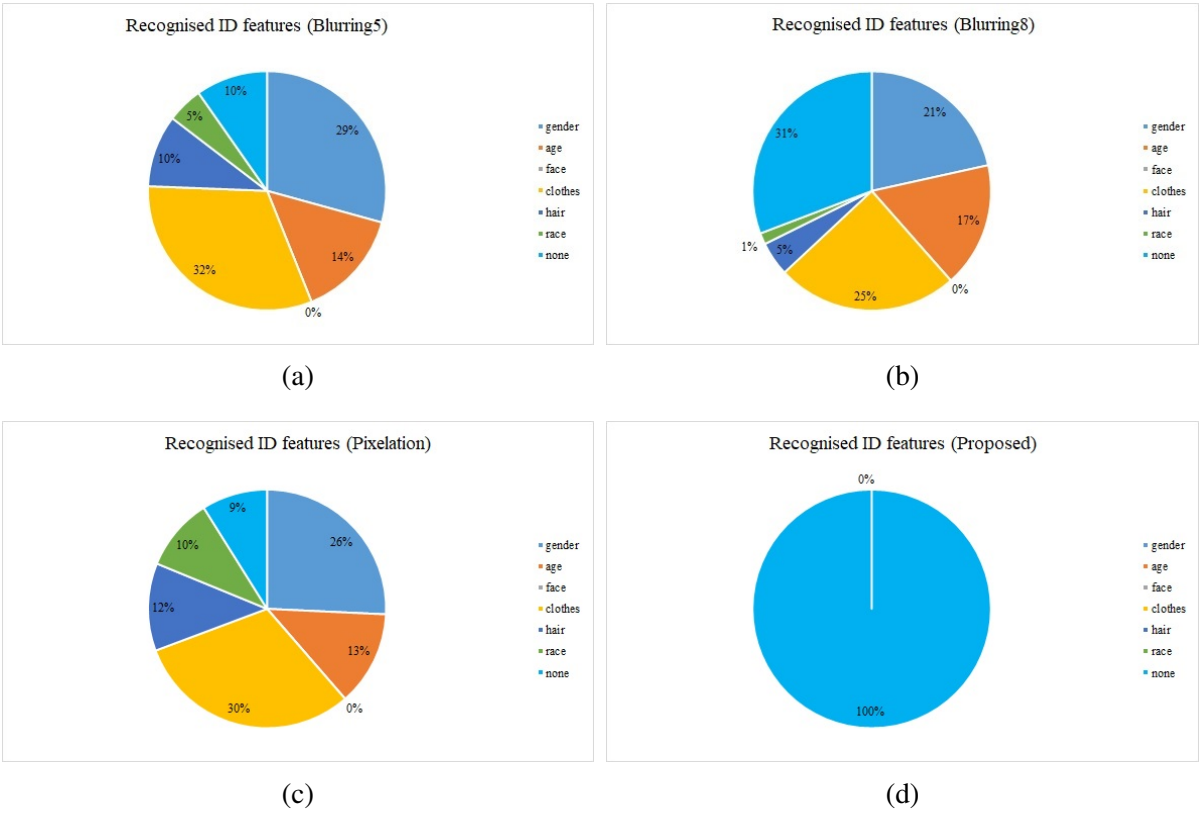


Figure 3.19: Percentages of recognising the appearance attributes in KTH dataset using each method: a comparison. The Silhouette and Binary methods are excluded because this dataset does not have a ground truth that can be used for getting the binary masks and forming the silhouettes.

3.4.4 Temporal salience-based anonymity objective evaluation

We adopted an objective evaluation, the magnitude of mean cross-correlation (MMCC), to evaluate the robustness of the proposed privacy protection approach and the current filtering algorithms w.r.t. the unmodified video frames from the perspective of similarity. Table 3.5 depicts the calculated MMCC for four filtering methods and the proposed method over all the sequences in DHA, KTH, Weizmann, and UIUC1 datasets. As we can see, the proposed method outperforms all filtering methods. This outperforming means that the temporal salience maps contain silhouettes that mostly have no similarity with RGB versions. By this, we can conclude that the proposed method achieves a high level of anonymity in terms of its ability to conceal the appearance details.

The objective evaluation results match the results of the subjective evaluation, which indicates that the proposed method achieves up to 100% of anonymity. Both the proposed method

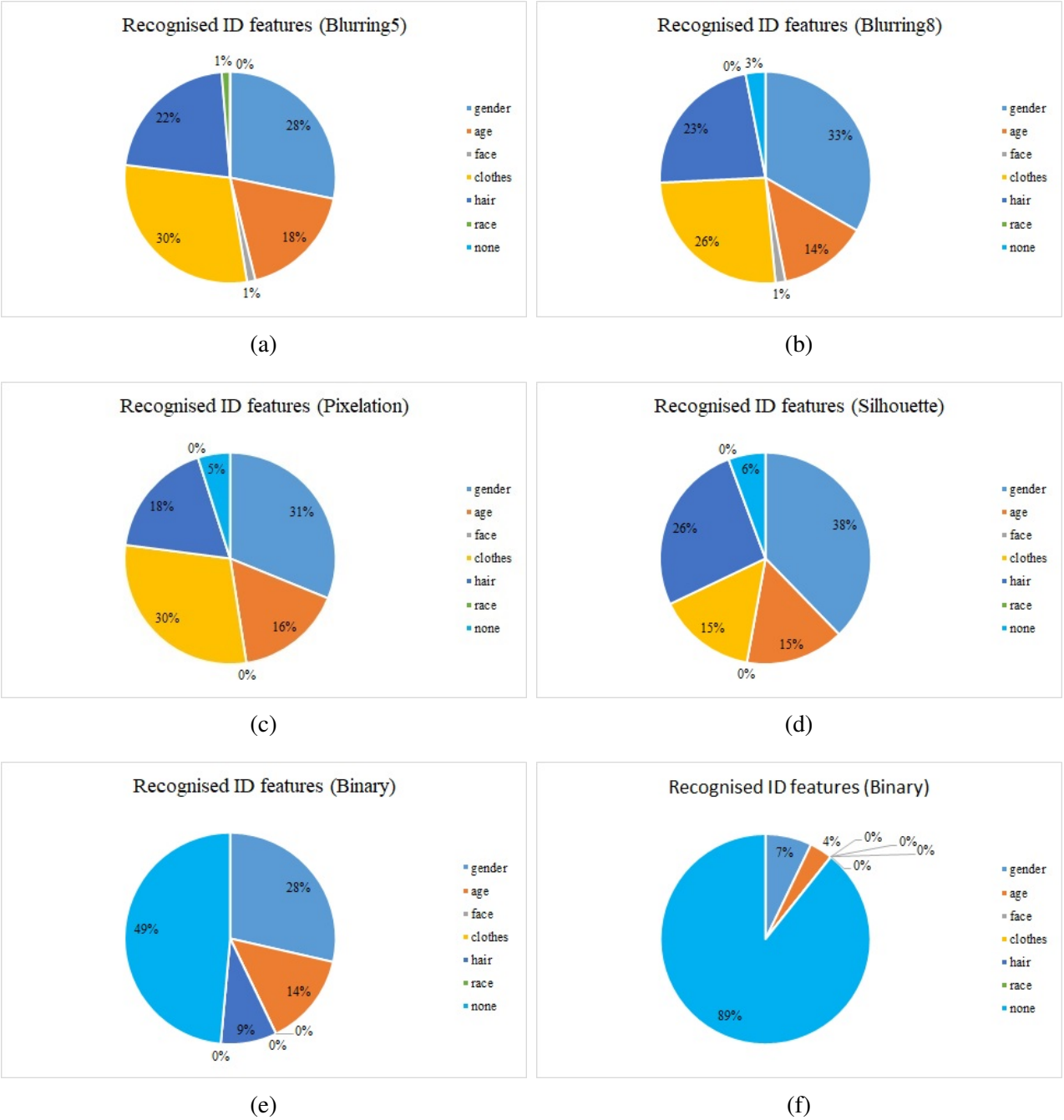


Figure 3.20: Percentages of recognising the appearance attributes in Weizmann dataset using each method: a comparison.

and the binary silhouette got the lowest scores of similarity. The MMCC values point out that our method achieves a high level of privacy preservation which is expected due to its strong anonymised action abstraction and modelling. The lowest MMCC means that the temporal saliency maps are considered masks which make the proposed anonymising method achieve a

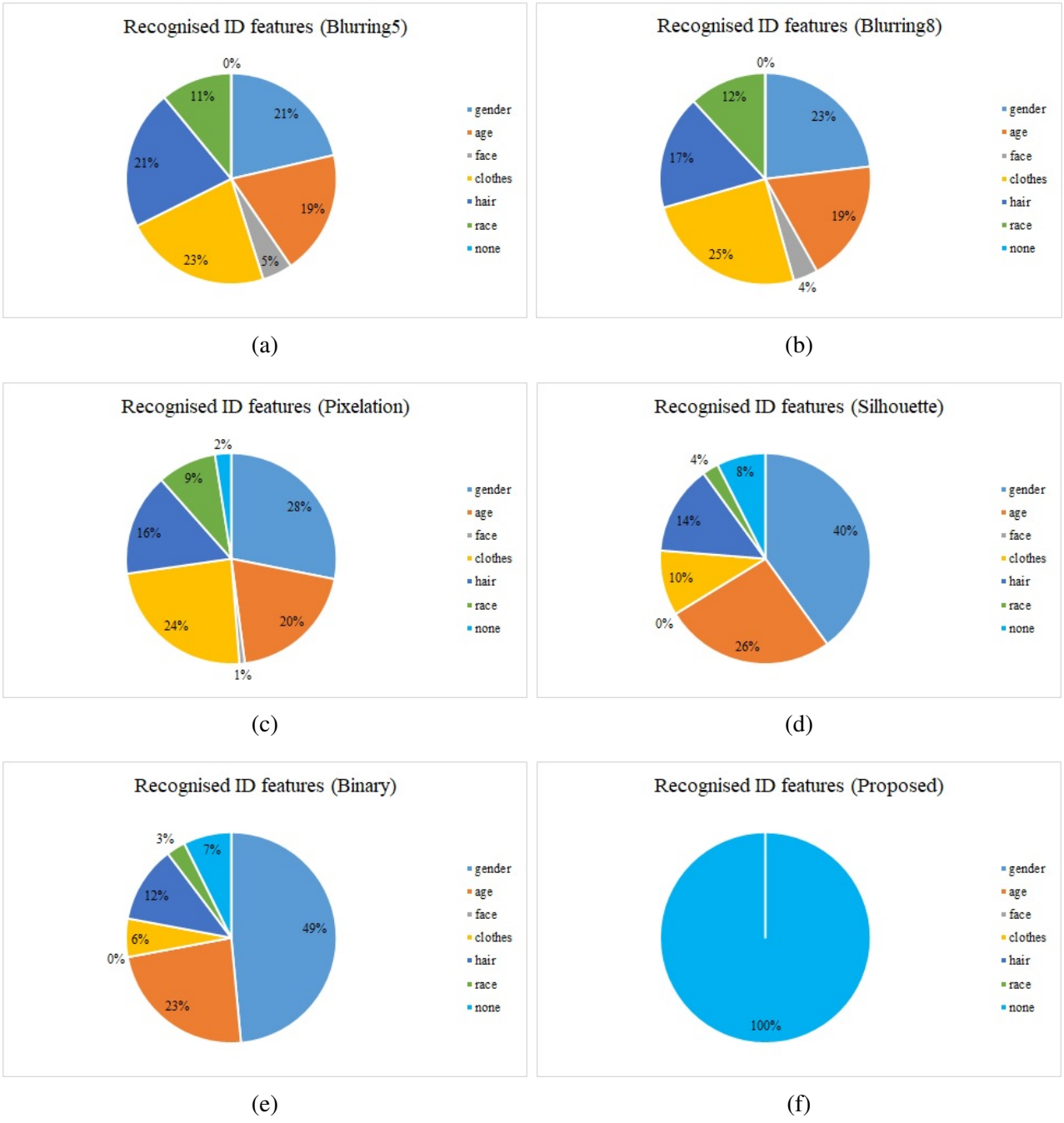


Figure 3.21: Percentages of recognising the appearance attributes in UIUC1 dataset using each method: a comparison.

high level of privacy protection similar to the binary mask. However, the difference between the binary mask and the saliency map is that the saliency map distributed the magnitudes in the silhouette region based on the dynamic changes acquired at each location. In contrast, the binary mask assigns the same magnitude overall locations. These variations will be observed in

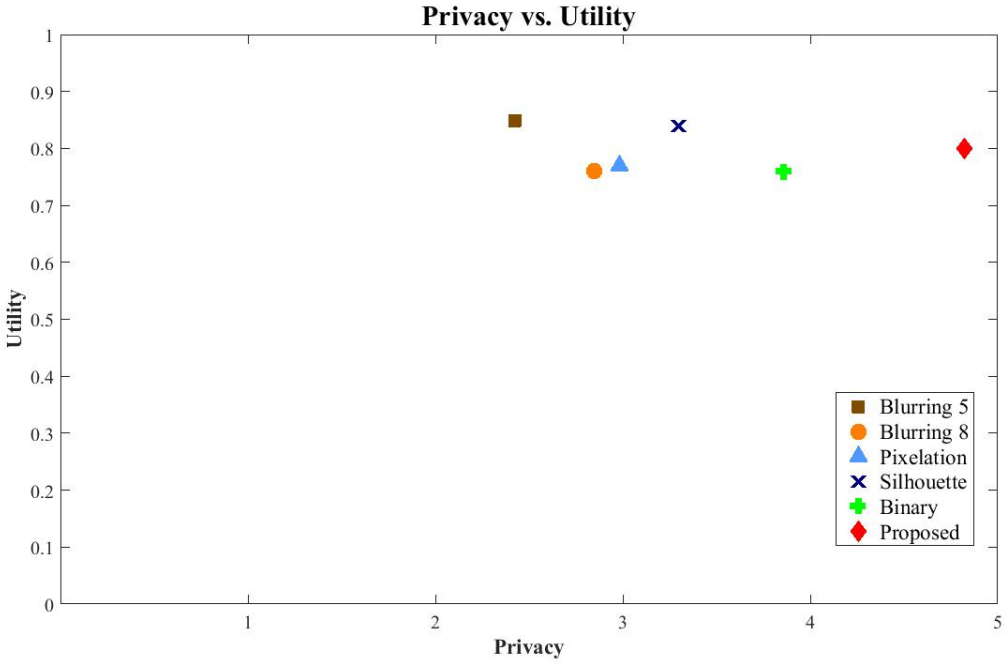


Figure 3.22: Privacy vs. utility of the vision-based anonymisation methods. The result of Silhouette and Binary results are collected from DHA, Weizmann, and UIUC1 datasets and excluded KTH dataset, since KTH dataset does not have binary masks that can be used to generate the Silhouette and the Binary models.

the human action recognition stage in the next chapter.

Back to Table 3.5, other methods of anonymisation in this table retain the intensity values of the origin and include them in the obfuscated version. These intensities maintain the appearance clues of the human leading to decreasing the privacy preservation. Therefore, depending on the spatial information to model the anonymity reduces the protection of privacy.

Table 3.5: Anonymising performance using the MMCC computed on the bounding box.

Method	DHA	KTH	Weizmann	UIUC1
Blurring $\sigma = 8$	0.90	0.8151	0.66	0.9653
Blurring $\sigma = 5$	0.94	0.8608	0.78	0.9836
Pixilation	0.74	0.8874	0.58	0.8795
Silhouette	0.68	N/A	0.84	0.8901
Binary Silhouette	0.48	N/A	0.75	0.691
Proposed Method	**0.02**	**0.3710**	**0.27**	**0.047**

3.4.5 Computational complexity of the temporal salience detection

All experiments in this chapter were implemented using Matlab R2017b on a PC with Intel processor, CPU@2.7GHz and RAM 8GB. The required time to obtain the temporal salience map is about 5.4 seconds. Table 3.6 explains the breakdown of the average times for each step of the proposed method. In our algorithm, we have three main steps to calculate the complexity. At the beginning, the frame difference is calculated, which is a simple absolute subtraction and this takes $\mathcal{O}(1)$. The next two steps are the 2DFFT and the spectral entropy for each 3×3 block size. Since these operations are executed based on the number of blocks extracted from each frame, the complexity for each one is $\mathcal{O}(M)$, where M is the number of blocks. Therefore, to complete processing all the blocks, we need $\mathcal{O}(2M)$. Thus, the total time complexity to compute each entropy map is $\mathcal{O}(2M) + \mathcal{O}(1)$. Since $M \gg 1$, the overall time complexity becomes $\mathcal{O}(M)$. This time complexity is calculated and explained in Table 3.1 whenever the execution time of the proposed anonymisation method is the best compared to the state of the art methods.

3.5 Concluding Remarks

In this chapter, we have presented a new temporal salience-based anonymisation method for privacy preservation in the application of video-based home monitoring. The proposed method relies on detecting the temporal change in the pixel intensities to model the daily human actions instead of filtering the spatial content of the visual data. The temporal salience-based silhouettes are used to achieve a high level of privacy obfuscation and present a useful content that can be utilised to identify the actions, *i.e.*, utility, regardless of the trade-off between the privacy

Table 3.6: The complexity of the proposed temporal salience estimation and obtaining HOG-S .

Step	Computational complexity
Frame difference	$\mathcal{O}(1)$
2DFFT	$\mathcal{O}(M)$
Spectral entropy	$\mathcal{O}(M)$
Total	$\mathcal{O}(2M) + \mathcal{O}(1)$

and utility. The proposed method has been evaluated using subjective and objective metrics to stand on the performance of the temporal salience-based privacy preservation. The results of the evaluation show that the proposed method achieves in a high level of privacy protection as well as concealment of the appearance clues up to 100% outperforming the existing filtering methods. The results of the subjective evaluation prove that the proposed method presents an informative abstract in the anonymity domain that can be used beyond the anonymisation to action recognition. The utility of the anonymity domain will be evaluated more from the perspective of the machine in the next chapter by extracting feature directly from the anonymity domain.

Chapter 4

Anonymised domain Human Action Recognition

This chapter proposes a new descriptor to represent the human action in the anonymised sequences. The proposed descriptor aims to emphasise the utility of exploring the anonymised content from the perspective of action recognition. This descriptor depends on exploring the anonymised video sequences for HAR. These obfuscated sequences have been already produced using the proposed temporal salience-based anonymisation method in Chapter 3.

4.1 Introduction

Vision-based Human Action Recognition (HAR) plays an important role in many applications, such as video surveillance [233], human-computer interaction [234], healthcare monitoring [235], assisted living [14, 16], smart homes [236] etc. Vision-based HAR is still a challenge due to different limitations, such as light conditions, occlusion and inherent redundant background. Although some of these problems can be overcome by acquiring a set of features to train a classifier leading to promising results, uncorrelated and lost information may be obtained during the feature extraction [237]. Therefore, the representation of action features is considered a critical stage affecting the performance of any action recognition system [238].

Low-level feature extraction is the conventional approach to represent the actions [237, 238, 133, 140, 135, 239, 189, 240, 241, 242, 243, 244, 118]. Accordingly, many works have been

proposed to represent the actions using either local visual descriptors [133, 140, 135, 239, 189, 240, 241, 242] or global visual descriptors [243, 244, 118]. Local descriptors, such as a Histogram of Oriented Gradient (HOG), Histogram of Optical Flow (HOF) and Motion Boundary Histogram (MBH), are widely used to represent the actions for HAR [245]. The global category of descriptors is also exploited to represent the actions in different scenarios [118, 243, 244]. However, vision-based algorithms suffer from several problems, such as the redundancy in successive frames, leading to extracting features that do not discriminate the actions accurately. One solution is to extract these features based on determining candidate local interest points to describe the actions [246]. However, these selected locations represent some parts of the human body, making the extracted features from these points adversely affect the accuracy of the classification.

Other problems, *e.g.*, privacy and utility, arise whenever vision-based HAR is exploited in the home monitoring scenarios. These systems struggle to find a compact methodology that can achieve both: good concealment for the privacy and exploring anonymity domain. The current solutions [50, 51, 224, 60], which have been explained in Chapter 3, attempt to find the trade-off between the privacy and utility. This trade-off means that increasing the privacy reduces opportunities to exploit that anonymised data and vice versa. Therefore, looking for a domain that achieves both the privacy and intelligibility using the same domain is considered the goal of every anonymity algorithm.

Recently, saliency estimation has attracted much attention in image and video processing [247, 3, 225]. The saliency estimation algorithms highlight the most important visual content, i.e. foreground, and attenuate others, i.e. background. The saliency also seems to be a useful tool for addressing the problems mentioned above of visual information that makes the saliency-based representation reliable and accurate for the feature learning applications. This representation has been exploited in [222] to improve the human action modelling by highlighting the most dominant foreground region and eliminating the background content to build an essential feature learning HAR system.

This chapter proposes a new approach that explores the temporal salience-based anonymity domain for HAR. The proposed method models the action without using high complexity motion estimation algorithms. Instead, the proposed method generates temporal saliency maps,

considering the spatial changes within successive frames. This modelling is followed by extracting HOG features leading to the histogram of oriented gradient of salience (HOG-S) features that are finally classified using the classifiers. The main contributions of this chapter are:

1. Exploring the temporal saliency domain instead of RGB domain for HAR [222]; *i.e.*, utility.

2. Proposing a salience-based descriptor to encode each action using the HOG of salience (HOG-S).

The rest of this chapter is organized as follows: Section 4.2 explains the related work in the field of HAR. Section 4.3 presents the proposed method to model the action and extract the features. Section 4.4 shows the experimental results and discussion followed by the conclusions in Section 4.5.

4.2 Related work

As we explained in chapter two, several works have been presented to represent the actions using the local dense trajectories representation, such as Histogram of Oriented Gradients (HOG) [245], due to its robustness [141]. The existing works on HOG-based HAR are categorised into two themes: 2D HOG [248, 249, 250] and 3D HOG [140, 251, 252] representations. In the first category, the dense features are extracted from a single image/frame to show the motion history. In the second category, a volumetric representation in space-time is exploited to represent the action. However, in both categories, redundant data, such as, the background, is exploited to extract features that represent the actions. This redundancy affects the discriminating power of the descriptor and increases storage requirements for this information and makes the complexity higher. Mainly, there is interest to address these problems based on determining candidate local interest points [246]; however, interested point-based learning has also many problems.

Recently, saliency estimation has attracted much attention in image and video processing [247, 3, 225]. The visual saliency estimation algorithms highlight the most important visual content, i.e. foreground, and attenuate others, i.e. background. The visual saliency offers a tool

for addressing the problems mentioned above of visual information [22, 253], and makes the saliency-based representation useful and accurate for the feature learning applications.

The usefulness of the saliency can be exploited to improve the human action modelling by highlighting the most dominant foreground region and eliminate the irrelevant background content to build an essential feature learning for HAR system. Besides, the saliency can improve the feature learning by guiding the descriptor toward the most important content in the scene, *i.e.*, the region of the action, and extract a reliable representation for the action to avoid the irrelevant content. The salience modelling also includes a variation in highlighting the details of the scene, which means that the discrimination is included in the content of the saliency map. Therefore, building feature learning based on the saliency leads to extract more discriminated features that can improve the feature learning and action recognition.

4.3 The proposed method

The flowchart of the proposed method is illustrated in Fig. 4.1. Let $C = \{s_i^F, l_i\}_{i=1}^V$ be the action dataset with V video sequences and L classes, where s_i is the ith RGB video sequence containing F frames and label $l \in L$. We aim to recognise V accurately into L classes using the new HOG-S descriptor. Our recognition system is composed of two main stages:

1. Temporal salience-based action anonymisation: The proposed anonymisation method has been done in Chapter 3.

2. Exploring the temporal saliency-based anonymised maps to extract the HOG-S . Based on modelling the actions, new efficient and more discriminating features are extracted from the anonymised maps to describe the action. This stage will be detailed in this chapter.

4.3.1 HOG-S feature extraction

The proposed description approach aims to construct a compact descriptor by excluding redundant information. Most current human action descriptors depend on the original RGB content to extract the features. The RGB domain has redundant regions, and the obtained features from

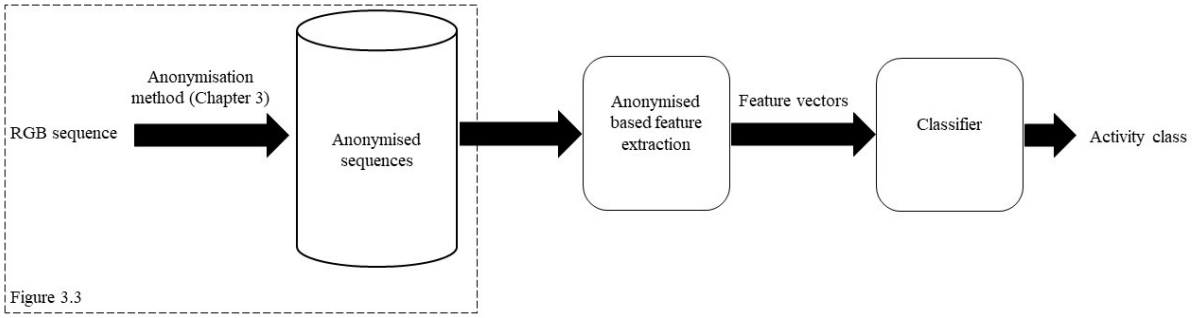


Figure 4.1: Proposed human action representation model.

these regions are included in the description and cause shortcomings in the accuracy of recognition. Therefore, guiding the descriptor toward the area included the action abstraction, *i.e.*, the saliency-based silhouette, leads to the outperforming in recognising the action.

HOG-S is a local descriptor exploring the anonymised sequences. This descriptor focuses on the salience region, avoiding unnecessary data to extract more informative content. By this, we guarantee that the redundant data is bound to be less contribution in describing the action. Our HOG-S is calculated from the saliency-based silhouette using the bounding box. All HOG-S vectors for the key-frame are calculated and collected to train the classifier.

The HOG-S focuses on the salience region, R_t, spanning in a rectangular bounding box of $K \times L$ pixels, from the silhouette in frame t. Major steps of our approach include HOG-S feature vector extraction from the bounding boxes, HOG-S feature vector processing and training a classifier as illustrated in the block diagram in Figure 4.2. We start by computing gradients, $\nabla R_t = (d_x, d_y)$ for each pixel in the region $R_{S_{\hat{\mathcal{E}}}}$, where d_x and d_y represent the horizontal and vertical components approximated by finite differences. The gradient magnitude, G_t, and the direction, θ_t, are computed as follows:

$$G_t = \sqrt{d_x^2 + d_y^2}, \tag{4.1}$$

$$\theta_t = arctan\left(\frac{d_y}{d_x}\right). \tag{4.2}$$

R_t is partitioned into $B_K \times B_L$ blocks, each containing $qm \times qm$ pixels. Then each block is further partitioned into $q \times q$ patches, with each patch containing $m \times m$ pixels. The gradient

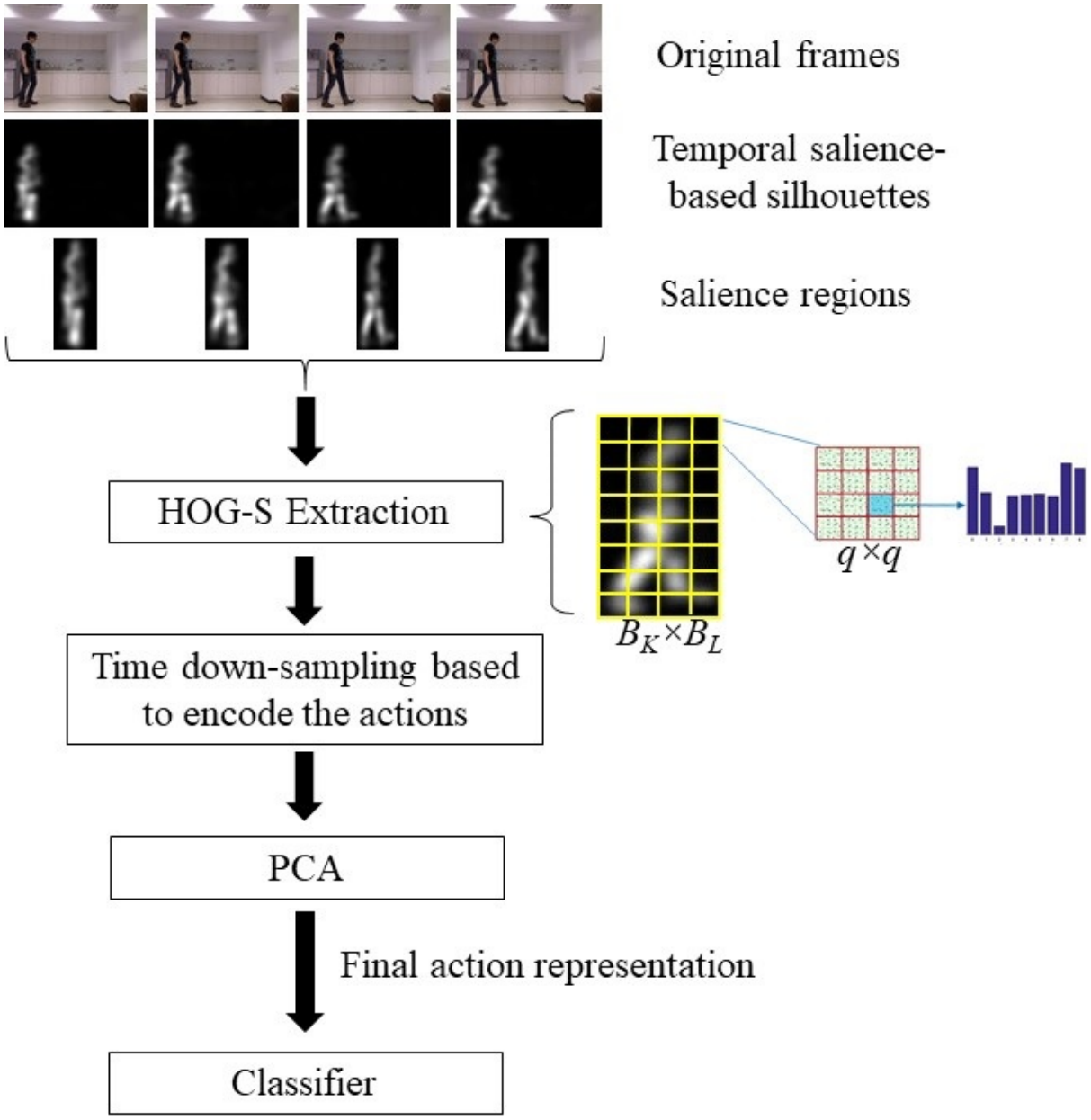


Figure 4.2: Proposed flowchart for HOG-S exploring the saliency-based silhouette of the human action.

magnitudes and the corresponding directions in each patch are formed into 9-bin histograms and all histograms are concatenated into a single feature vector, $\vec{v}$, of length $9q_2B_KB_L$. This is followed by normalizing the vector as follows:

$$\hat{v}_t = \frac{\vec{v}_t}{\|\vec{v}_t\|_2^2}. \tag{4.3}$$

However, just considering individual $\hat{v}_t$ for individual frames cannot perfectly marginalise among features from other frames in accordance with the variations inside the action itself and similarities among other actions. This is addressed by considering the accumulated temporal changes to the feature vectors, $\hat{\mathbf{V}} = \{\hat{v}_0, \hat{v}_1, \hat{v}_2, \cdots, \hat{v}_t\}$ up to frame t to compute the final feature vector, $\tilde{v}_t$, at the time instant, t, as follows:

$$\tilde{v}_t = \left| \sum_{k=0}^{t-1} \hat{v}_{t-2k} - \sum_{k=0}^{t-1} \hat{v}_{t-(2k+1)} \right|, \tag{4.4}$$

where $t - k > 0$.

Eq. (4.4) improves the HOG-S features by increasing the discrimination among the actions. Besides, the approach can be further improved by applying the principal component analysis (PCA) on $\tilde{v}_t$ to satisfy the following two objectives:

1. To reduce the length of the HOG-S descriptor.

2. To improve the discrimination of our HOG-S descriptor.

In the section of the performance evaluation, we verify the improvement of the temporal salience-based feature extraction algorithm for HAR by exploring the region of the action; *i.e.*, the temporal salience silhouette, and avoiding the redundancy.

4.4 Performance evaluation

In this section, the proposed HOG-S descriptor is evaluated on six publicly available action recognition datasets, namely Weizmann[118], KTH[4], DHA [212], UIUC1 [119], UCF sports [183, 180], and HMDB51 [205], to verify the utility of the proposed method in Chapter 3.

4.4.1 Datasets

In the following, we will briefly explain the datasets that are used in the experiments. These datasets have been already explained in Chapter 2, however, we will explain them in terms of the notations specified in Section 4.3.

1. Weizmann dataset was first proposed by Blank *et al.*in [114] and contains 93 video sequences. These sequences have low-resolution sequences of 144×180 with a frame rate of 50 frames per second (fps). The dataset shows nine actors, each of them performing $L = 10$ different actions, i.e. bend, run, walk, skip, jack, jump, pjump, side, one hand wave and two hands wave. This dataset is widely used in the applications of action recognition.

2. KTH dataset was proposed by Schuldt *et al.*[4] and showing $L = 6$ actions, *i.e.*, boxing, handwaving, handclapping, jogging, running and walking. The dataset was acquired over homogeneous backgrounds with a static camera recording 25 fps frame rate. Twenty-five different subjects are performing the actions mentioned above in four different scenarios. Each sequence has a resolution of 160×120 with an average of 4 seconds length.

3. UIUC1 is an indoor dataset [119] includes $V = 532$ sequences showing $L = 14$ human actions, i.e., walking, running, jumping, waving, jumping jacks, clapping, jump from situp, raise one hand, stretching out, turning, sitting to standing, crawling, pushing up and standing to sit, captured by a static camera. These 14 actions are performed by eight actors, where each actor does the same action several times. The sequences came with a resolution of 1024×768 and 15 fps frame rate.

4. Depth-included Human Action (DHA) dataset was suggested by Lin *et al.*[212]. The dataset consists of $V = 532$ sequences compromising $L = 23$ action categories performed by 21 subjects (12 males and 9 females). It is recorded using a static Kinect camera in three different scenes with 480×640 resolution. The RGB versions of sequences are used in the experiments.

5. Human Motion Database (HMDB) [205] is one of the largest datasets used to recognise the human motion contains $V = 6849$ clips distributed in $L = 51$ action classes; each video has $\sim$ 20–1000 frames. The actions categories of this dataset can be grouped into five types based on body movements. This dataset is considered a challenge due to including sequences collected from the Internet and YouTube; therefore, this dataset is a real-world video sequences collection.

6. University of Central Florida (UCF) sports dataset [180, 183] is a set of action sequences collected from the broadcast channels from various sports contains $V = 150$ sequences with the resolution of 720×480 over $L = 10$ classes. The action features a wide range of scenes and viewpoints. This dataset has been used for the application of computer vision; such as action recognition and action localization.

4.4.2 Experiment Setup

In the experiments, the considered number of user-defined thresholds equals to 7. All silhouette maps are resized to the resolution of 256×256 to apply the same parameters on all datasets. We adopt a bounding box approach with 168×72 resolution to crop the temporal salience region of the silhouette. We found that the patch size 4×4 with $P = 16$, resulting in 144-dimensional descriptors for each block achieves the best results. The final dimension of HOG-S descriptor for each frame in all datasets is 23040.

4.4.3 Comparison Results

Experiments reported in this section perform an objective evaluation of the visually anonymised sequences from the proposed anonymisation method from the machine perception. We report its performance in the datasets mentioned above, *i.e.*, Weizmann, KTH, UIUC1, DHA, HMDB51, and UCF sports using both KNN and SVM classifiers with five-fold cross-validation and compare with the existing methods.

We will start by showing the confusion matrix and the table of comparison with the existing work on each dataset, separately. The experimental results on Weizmann dataset are illustrated in Table 4.1. Our descriptor achieves average recognition accuracies of 99.46% and 99.66% using QSVM and KNN, respectively, outperforming most state of the art methods. Figures 4.3 and 4.4 show the confusion matrices of QSVM and KNN classifiers, respectively, based on HOG-S . The results show that 30% of the actions are recognised with 100% accuracy using QSVM and 50% are recognised with 100% accuracy using KNN. This performance proves that the proposed HOG-S accurately recognises between the actions, though there are similarities between the actions; such as between one hand wave and two hand wave actions. Besides, we

Table 4.1: Recognition accuracy (%) of the proposed HOG-S and the state of the art on Weizmann dataset: a comparison.

Method	Accuracy %
Klaser *et al.*(2008)[140]	84.3
Weinland and Boyer (2008)[173]	93.6
Ta *et al.*(2010)[174]	94.5
Xie *et al.*(2011)[175]	95.60
Wu and Shao (2013)[170]	97.98
Zhang *et al.*(2015)[176]	96.3
Zeng *et al.*(2018)[177]	98.77
Xu *et al.*(2017)[178]	99.1
Rodriguez *et al.*(2017)[179]	98.9
Proposed using QSVM	**99.46**
Proposed using KNN	**99.66**

observe that KNN fully distinguishes the frames of run action despite the existence of similarity of action patterns with the walk action. The reason is that the new HOG-S descriptor is efficiency discriminate between the actions, even though they have the same patterns in some cases.

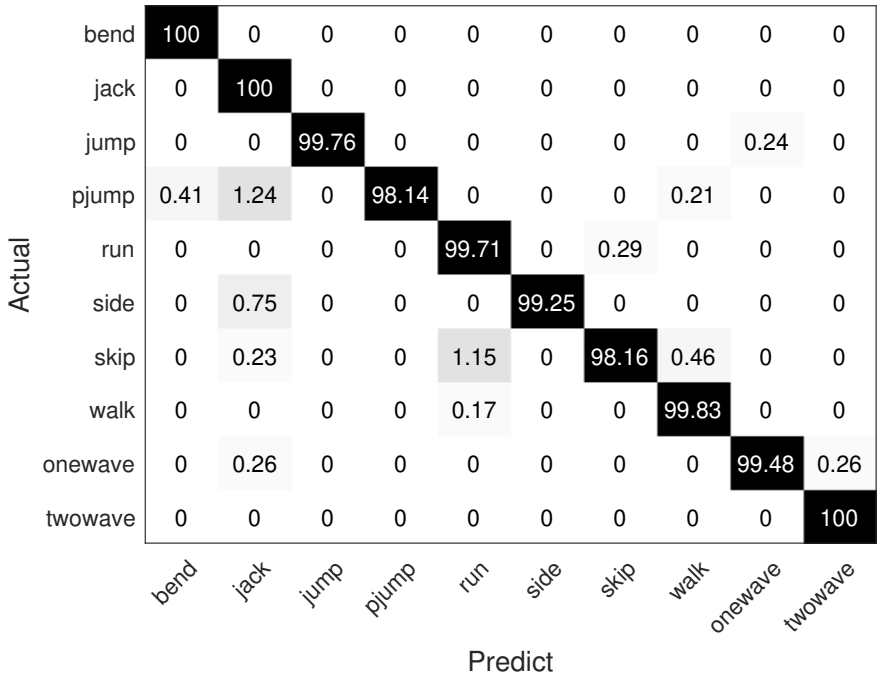


Figure 4.3: The confusion matrix of classification all views using QSVM on Weizmann (Overall accuracy: 99.46%).

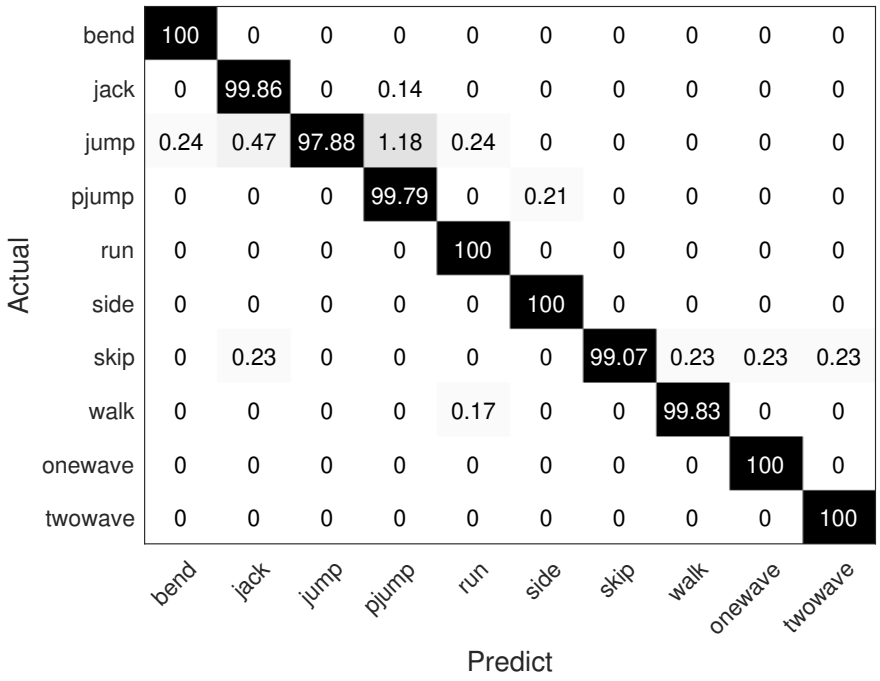


Figure 4.4: The confusion matrix of classification all views using KNN on Weizmann (Overall accuracy: 99.66%).

This good discrimination is due to applying the proposed temporal salience method of modelling the human body parts. Variation in actions modelling is formed based on the motion pattern of each action that can make it easy to distinguish between those actions since the motion patterns are fundamentally different from action to action.

The next dataset in our experiments plan is KTH. The comparison results of the accuracy rates are shown in Table 4.2. We can see that the proposed descriptor outperforms the existing work on this dataset by achieving average recognition accuracies of 98.53% and 99.06% using QSVM and KNN classifiers, respectively.

The accuracy results in Table 4.2 show that exploring the temporal salience of the actions is useful in a multi-view HAR scenario because the saliency highlights the same regions regardless of the view orientation of the camera, *i.e.*, view-invariant. Thus, the descriptor will target the same salience content that leads to calculating a description with high similarity to those from other camera views. This outperforming is due to the proposed method of modelling the action as a temporal salience silhouette. The corresponding confusion matrices of recognising the actions in KTH dataset using QSVM and KNN classifiers are shown in Figures 4.5 and 4.6,

respectively.

The third experiment has been applied on another indoor dataset; i.e., UIUC1. The results of the accuracy and the comparison have been shown in Table 4.3. Our proposed method achieves an overall recognition accuracy of 99.15% and 99.06% using KNN and QSVM classifiers, respectively. Regarding the accuracies in Table 4.3, the proposed descriptor outperforms improvement by 0.25% compared to state of the art on this dataset. The confusion matrices of

Table 4.2: Recognition accuracy ($\%$) of the proposed HOG-S and the state of the art on KTH dataset: a comparison.

Method	**Accuracy %**
Ikizler-Cinbis and Sclaroff (2012) [169]	81.17
Wu and Shao (2013)[170]	83.30
Liu *et al.*(2013) [128]	94.8
Veeriah *et al.*(2015)[171]	93.96
Yadav *et al.*(2016)[172]	98.20
Shi *et al.*(2017) [126]	96.8
Proposed using QSVM	**98.53**
Proposed using KNN	**99.06**

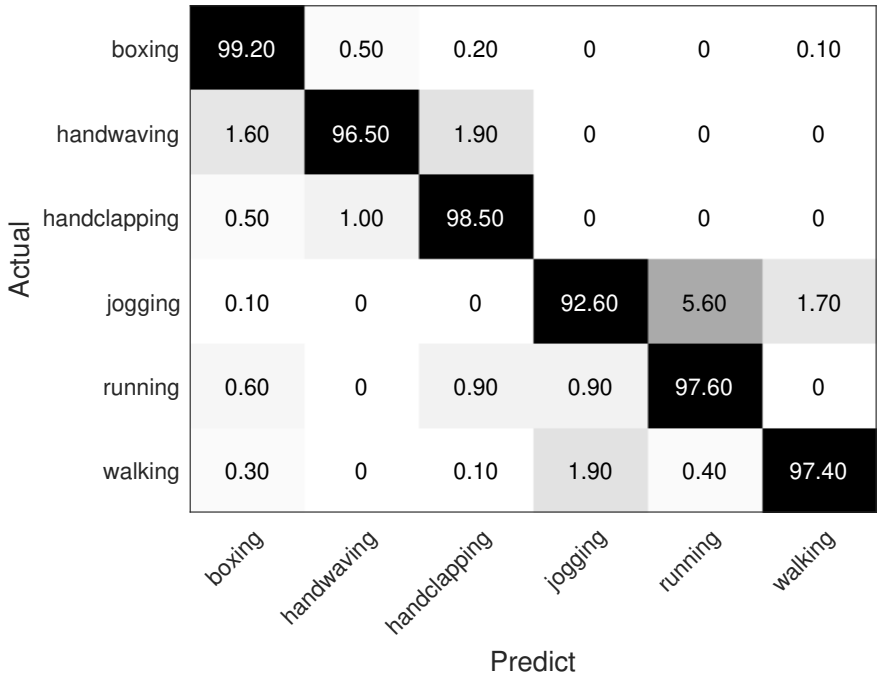


Figure 4.5: The confusion matrix of classification all views using QSVM on KTH (Overall accuracy: 98.53%).

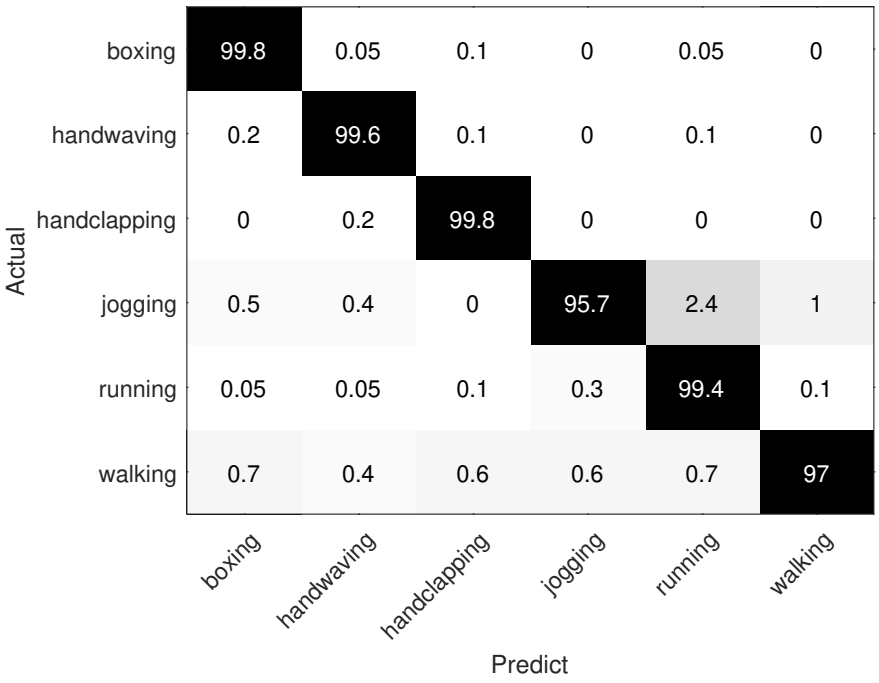


Figure 4.6: The confusion matrix of classification all views using KNN on KTH (Overall accuracy: 99.06%).

KNN and QSVM classifiers are depicted in Figures 4.7 and 4.8, respectively. In Figure 4.8, we can see that the classifier recognises the jumping action with 100% of accuracy despite the similarity between this action and other actions in the dataset. Besides, 79% of activities have been recognised with $> 99\%$ accuracy. These results prove the outperforming of the proposed method compared to state of the art. The results of a perfect discrimination are also obtained in Figure 4.7.

The fourth dataset in this scenario is DHA. This dataset contains twenty-three controlled

Table 4.3: Recognition accuracy ($\%$) of the proposed HOG-S and the state of the art on UIUC1 dataset: a comparison.

Method	**Accuracy %**
Parikh and Grauman (2011)[188]	93.4
Wang *et al.*(2013) [189]	98.4
Zhang *et al.*(2015) [190]	98.87
Shan *et al.*(2015) [191]	98.9
Proposed using QSVM	**99.06**
Proposed using KNN	**99.15**

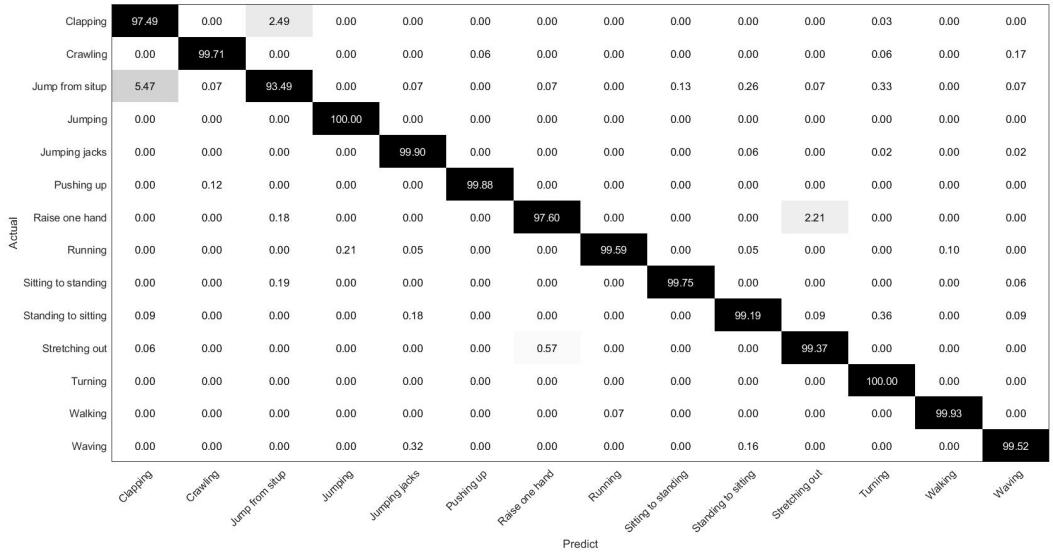


Figure 4.7: Confusion matrix of KNN classifier on UIUC1 dataset (Overall accuracy: 99.15%).

scene actions which are used to evaluate the utility of the temporal saliency maps for HAR. The results of recognising the actions in DHA dataset are shown in Table 4.4, Figure 4.9 and Figure 4.10, respectively. Though DHA dataset includes several actions with high similarity, our proposed method discriminates them accurately and outperforms the existing methods to achieve approximately 3% improvement, as can be seen Table 4.4. This outperforming is because our

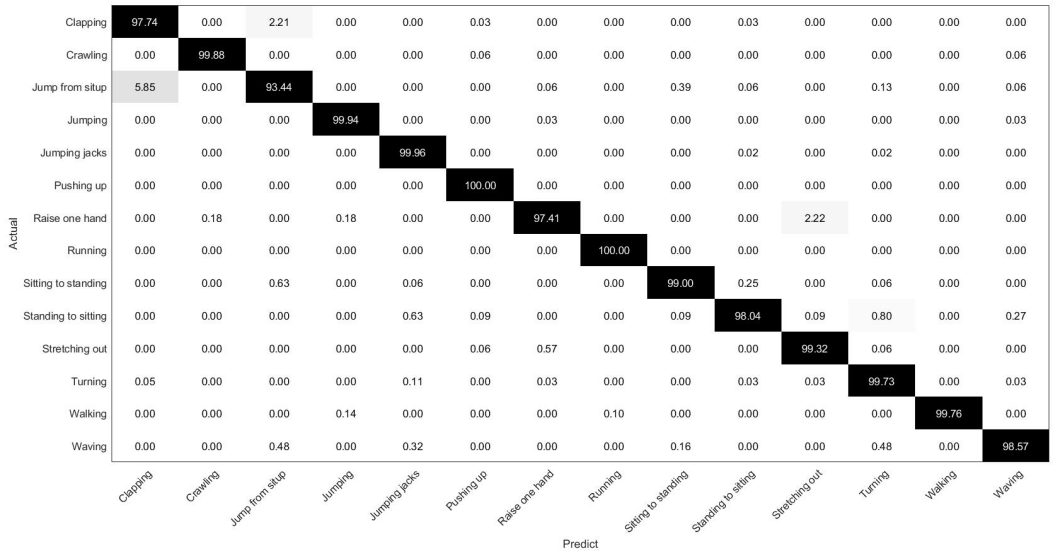


Figure 4.8: Confusion matrix of QSVM classifier on UIUC1 dataset (Overall accuracy: 99.06%).

Table 4.4: Recognition accuracy (%) of the proposed HOG-S and the state of the art for DHA: a comparison.

Method	Accuracy (%)
Liu *et al.*(2012) [212]	87
Yang *et al.*(2012) [213]	86.5
Gao *et al.*(2015)[214]	95
Liu *et al.*(2017)[215]	95.45
Zhang *et al.*(2017)[216]	96.69
Liu *et al.*(2018)[217]	95.44
Proposed using QSVM	**97.98**
Proposed using KNN	**99.59**

proposed method models the action silhouette differently for each action. Figures 4.9 and 4.10 also show the performance of the proposed method. The confusion matrix of KNN classifier shows that 8 out of 23; *i.e.*, 34%, of actions have been fully recognised based on exploring the silhouettes obtained by the proposed modelling method. Besides, the proposed representation leads to full recognition between the actions with high similarities, such as side-box and side-clap actions. The perfect discrimination among the actions indicates the accuracy of modelling the actions based on the proposed temporal salience method, and the reliability of applying the proposed HOG-S .

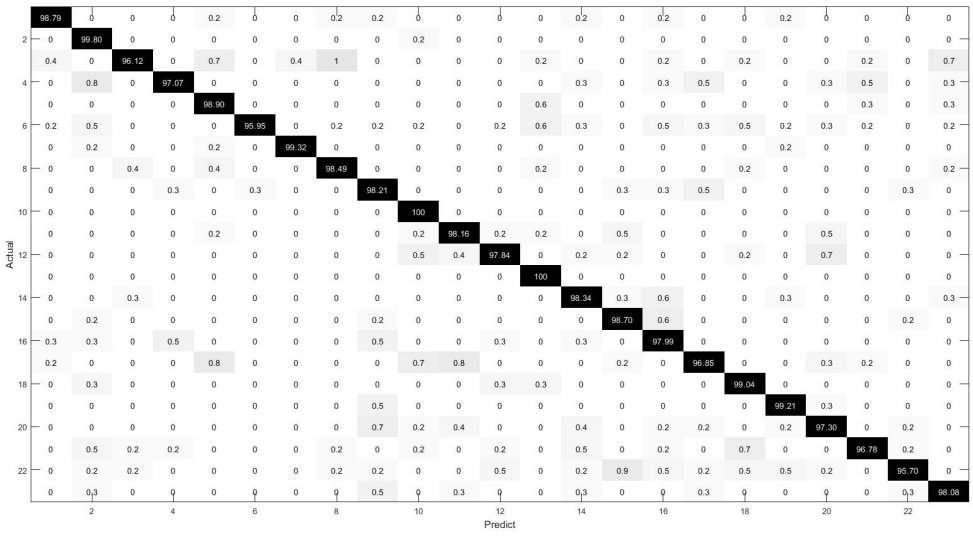


Figure 4.9: Confusion matrix of QSVM classifier on DHA dataset (Overall accuracy: 97.98%).

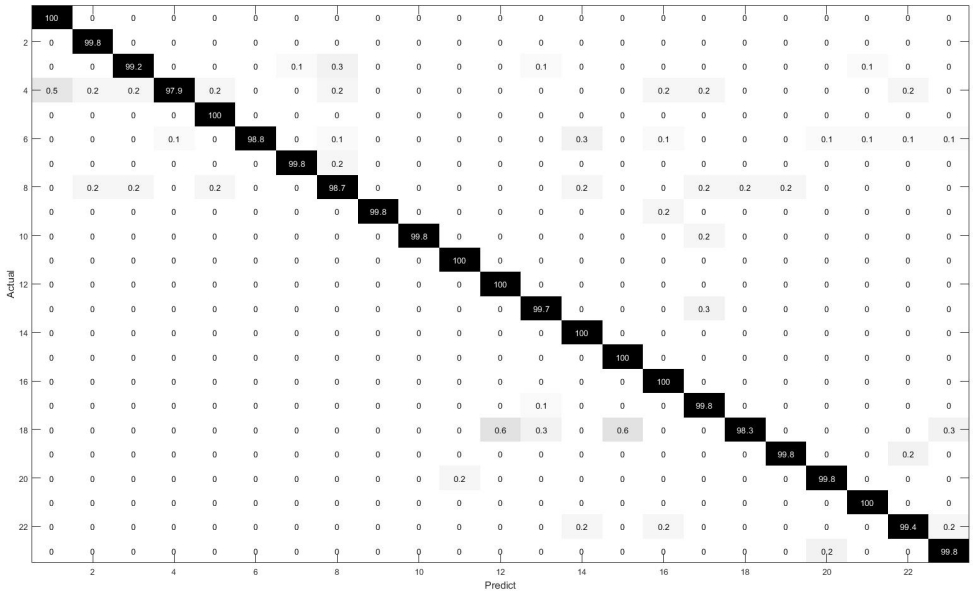


Figure 4.10: Confusion matrix of KNN classifier on DHA dataset (Overall accuracy: 99.59%).

Fifth, in this experiment, the proposed descriptor has been tested on more natural wild sequences collected from the Internet using HMDB51 dataset. The results of recognising the actions in this dataset are shown in Table 4.5 and Figure 4.11, respectively. Our new descriptor outperforms the existing work on this dataset and achieves approximately 16.55% improvement (see Table 4.5). This particular outperforming is due to considering the variation in collecting the sequences for each action class and the number of sequences included in each class. The confusion matrix of KNN classifier shows that 21 out of 51; *i.e.*, 41%, of actions have been fully recognised based on exploring the silhouettes obtained by the proposed modelling method. The accurate discrimination between the actions in HMDB51 dataset also proves the superiority of the proposed modelling method and the proposed HOG-S descriptor. In this part of the experiments, we have excluded the results of QSVM classifier because the specification of our pc hardware prevents us from completing the results since the size of this dataset set is huge compared to other datasets in this research. Therefore, we presented the results of KNN classifier only for HMDB51 dataset.

In the sixth experiment, the proposed descriptor has applied on more natural wild sequences collected from the Internet, *i.e.*, UCF sports action dataset. Applying the HOG-S to extract the

Table 4.5: Recognition accuracy (%) of the proposed HOG-S and the state of the art for HMDB51: a comparison.

Method	Accuracy (%)
Tran *et al.*(2015) [206]	51.6
Tran *et al.*(2017) [207]	54.9
Girdhar *et al.*(2017) [208]	66.9
Carreira and Zisserman (2017) [209]	80.9
Choutas *et al.*(2018) [210]	80.9
Wang *et al.*(2018) [211]	82.48
Proposed using KNN	**99.03**

features improves the accuracy of state of the art by 3.49%, as we observe in Table 4.6. Using the HOG-S with KNN classifier results in full recognition of around 77% of the actions (see Figure 4.12). The confusion matrix of using QSVM in Figure 4.13 also shows that 54% of the action are recognised with 100%.

The useful abstract created by the temporal salience and targeting this salience area for extracting the features has been entirely affected by the discrimination among the actions, leading to improving the accuracy of recognition. The results of this chapter explain that the anonymity domain based on the proposed method has a high level of utility with maintaining a high-level of anonymisation. Both privacy and utility are achieved together using the same approach without using other algorithms.

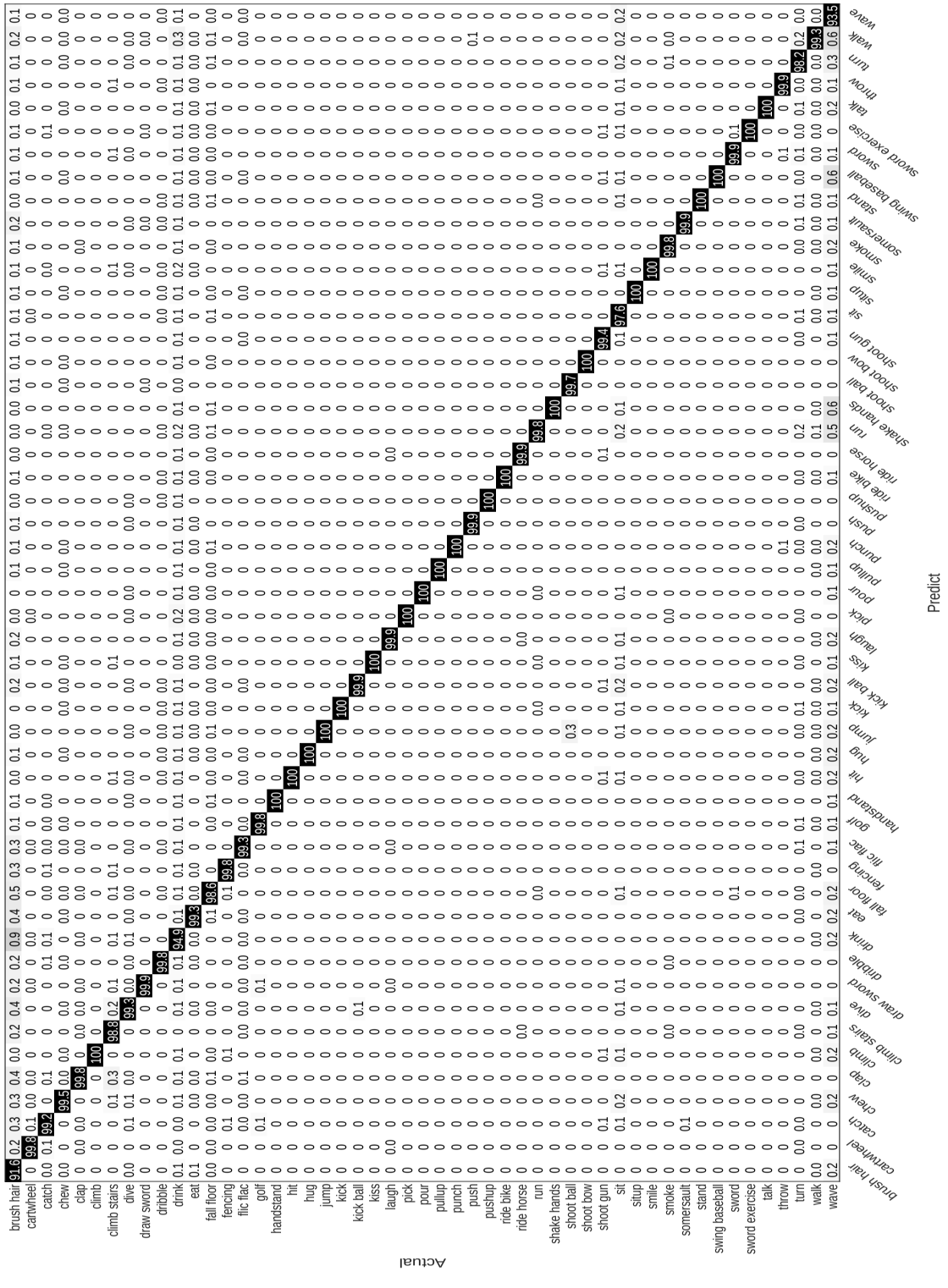


Figure 4.11: Confusion matrix of KNN classifier on HMDB51 sport dataset (Overall accuracy: 99.03%).

Table 4.6: Recognition accuracy (%) of the proposed HOG-S and the state of the art for UCF sports: a comparison.

Method	Accuracy (%)
O'Hara and Draper [181] (2012)	91.3
Shao *et al.*[182] (2013)	93.4
Soomro *et al.*[183] (2014)	92.67
Wang *et al.*[184] (2017)	93.6
Ghodrati *et al.*[185] (2017)	95.7
Wang *et al.*[186] (2018)	91.89
Siddiqi *et al.*[187] (2019)	96.22
Proposed using QSVM	**98.15**
Proposed using KNN	**99.71**

4.4.4 PCA-based HOG-S improvement

The proposed HOG-S descriptor is reliable to deal with different scenarios to satisfy the utility in HAR. However, this descriptor can be improved by reducing the length of the feature vector to optimise the complexity of training. This improvement can be made by applying the principal component analysis (PCA) and by choosing the optimal amount of components that can achieve

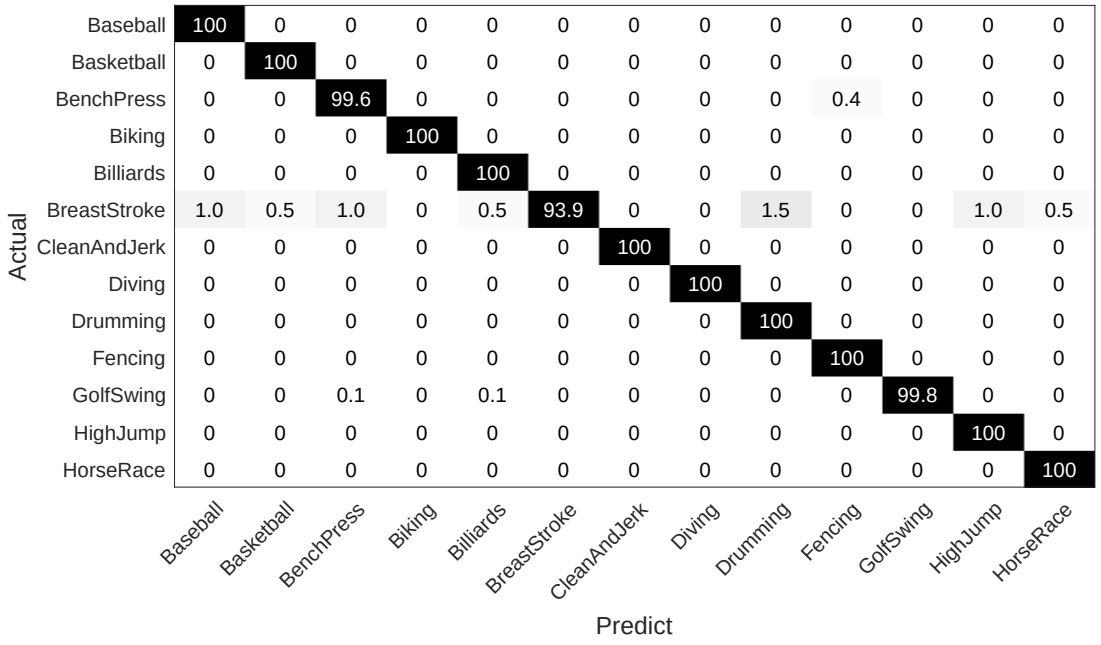


Figure 4.12: Confusion matrix of KNN classifier on UCF sports dataset (Overall accuracy: 99.71%).

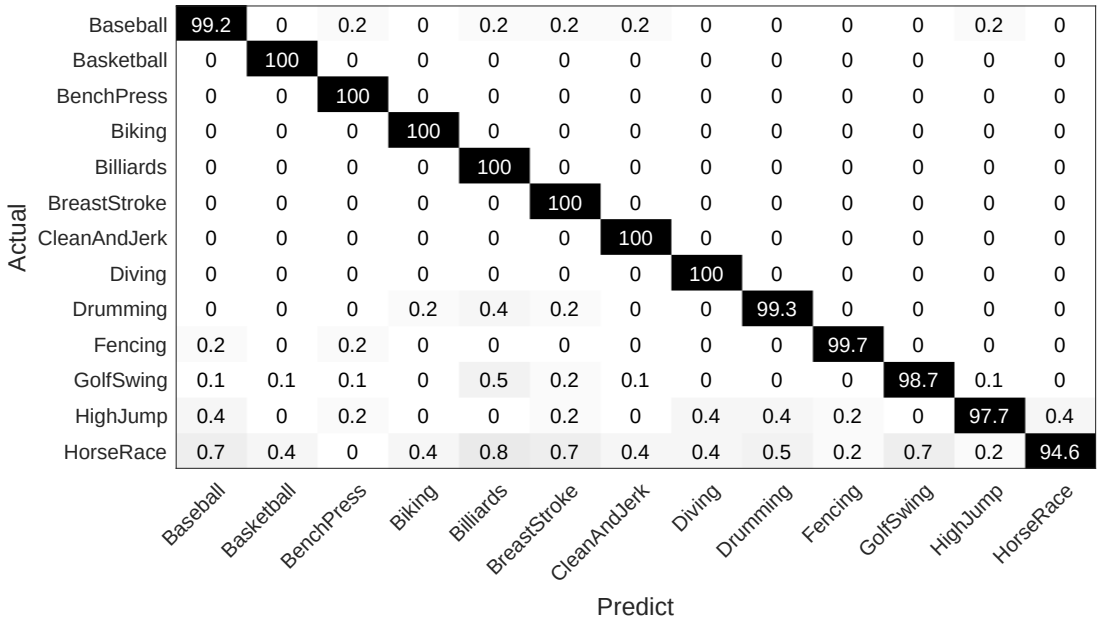


Figure 4.13: Confusion matrix of QSVM classifier on UCF sports dataset (Overall accuracy: 98.15%).

the best accuracy. Besides, reducing the dimensionality of the HOG-S , applying the PCA also increases the discrimination among the actions leading to improving the accuracy rates of recognition. For instance, the experiments on UCF sports dataset shows that the accuracy rate of recognition has been increased after applying PCA on HOG-S by 3.67% compared to state of the art. Table 4.7 shows more results of improvements before and after applying PCA on the extracted feature vector.

The number of PCA components that are selected to achieve the improvement in each dataset is depicted in Table 4.8. These PCA components are also explained in Figures 4.14 to 4.19 for Weizmann, KTH, DHA, UIUC1, UCF sports, and HMDB51, respectively. The selected number of components is highlighted with an orange colour in these figures. We notice that KNN classifier needs PCA components less than QSVM since QSVM classifier considers more information when classifying.

Table 4.7: Recognition accuracy (%) and the percentage of improvement of the proposed HOG-S before and after applying PCA on six datasets.

Method	Accuracy (%)					
	Weizmann	DHA	KTH	UIUC1	UCF sports	HMDB51
Proposed using HOG-S +QSVM	99.46	97.98	98.53	99.06	98.15	–
Proposed using HOG-S +KNN	99.66	99.59	99.06	99.15	99.71	99.03
Proposed using HOG-S +PCA+QSVM	**99.81**	**99.18**	**99.87**	**99.73**	**99.23**	–
Proposed using HOG-S +PCA+KNN	**99.74**	**99.73**	**99.94**	**99.72**	**99.89**	**99.19**
Percentage of improvement before PCA	N/A	2.9%	2.26%	0.25%	3.49%	16.55%
Percentage of improvement using PCA components	N/A	3.04%	3.14%	0.83%	3.67%	16.71%

4.4.5 Privacy vs. action recognition: a machine perspective

To sum up the results of the utility, *i.e.*, action recognition, from the perspective of the machine, we compare the response of the classification versus privacy. In this context, we used the results of privacy subjective evaluation from Chapter 3 and also included the results of the KNN classifier as the machine's utility to verify the comparison. This comparison is shown in Figure 4.20, where the methods of blurring, pixelation, silhouette, and binary mask are used to evaluate the utility of the anonymity domain from the viewpoint of the machine compared to the proposed temporal salience anonymisation method. The results in Figure 4.20 are collected from DHA, Weizmann, and UIUC1 datasets, and we used these datasets as samples in this comparison. Table 4.9 shows the accuracy rates for the anonymisation methods that are used in Figure 4.20.

We notice that the proposed method outperforms the existing methods and achieves the

Table 4.8: The number of PCA components used to improve the accuracies for five datasets.

Dataset	PCA components	
	KNN	QSVM
Weizmann	400	900
KTH	700	1300
DHA	300	1400
UIUC1	500	1700
UCF sports	400	400
HMDB51	900	–

Table 4.9: Accurate rates (%) of state of the are anonymisation methods and the proposed method for HAR.

Method	DHA	Weizmann	UIUC1	Average
Blurring5	94.62	94.05	98.94	95.87
Blurring8	94.51	91.15	99.18	94.95
Silhouette	93.9	89.36	98.99	94.08
Binary	91.97	93.61	98.19	94.59
Pixelation	79.35	69.16	93.23	80.58
Proposed	**99.39**	**99.64**	**99.15**	**99. 39**

highest accuracy rate for HAR. This result is because the proposed temporal salience-based anonymisation maintains the quality of the data and provides a useful action model that can be exploited beyond the anonymisation. The worst accuracy rate is caused by pixelation method which means that this method does not maintain data quality. The results in Figure 4.20 are comparable to those in Figure 3.22. We observe that the proposed method still has the best performance compared to the state of the art.

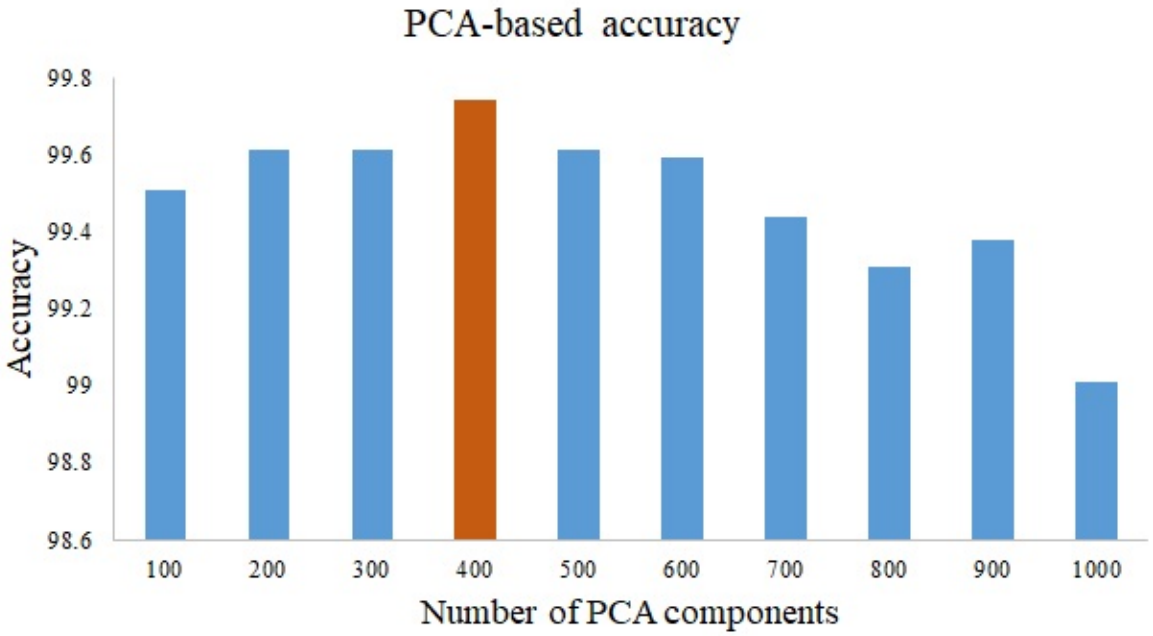
4.5 Concluding Remarks

In this chapter, a new descriptor exploring the temporal saliency maps for HAR has been proposed; *i.e.*, histogram of oriented gradient of salience (HOG-S). HOG-S exploits the temporal saliency maps to extract discriminated features. The salience region guides the descriptor toward the region of the action to capture the representation of the action leading to avoid the redundancy and the wasted time required to process the background contents. Two classifiers, i.e. QSVM and KNN, have been utilised for training and testing the temporal saliency maps. Several experiments using six standard datasets for HAR are conducted to evaluate the proposed descriptor. The presented results of the accuracy show the superiority of the proposed descriptor over the existing methods. These accuracy rates show that the proposed descriptor can efficiency discriminate the actions despite including similarities between them. This outperforming of HOG-S is due to the proposed method of modelling the temporal salience silhouettes, which contributes to calculating a useful description.

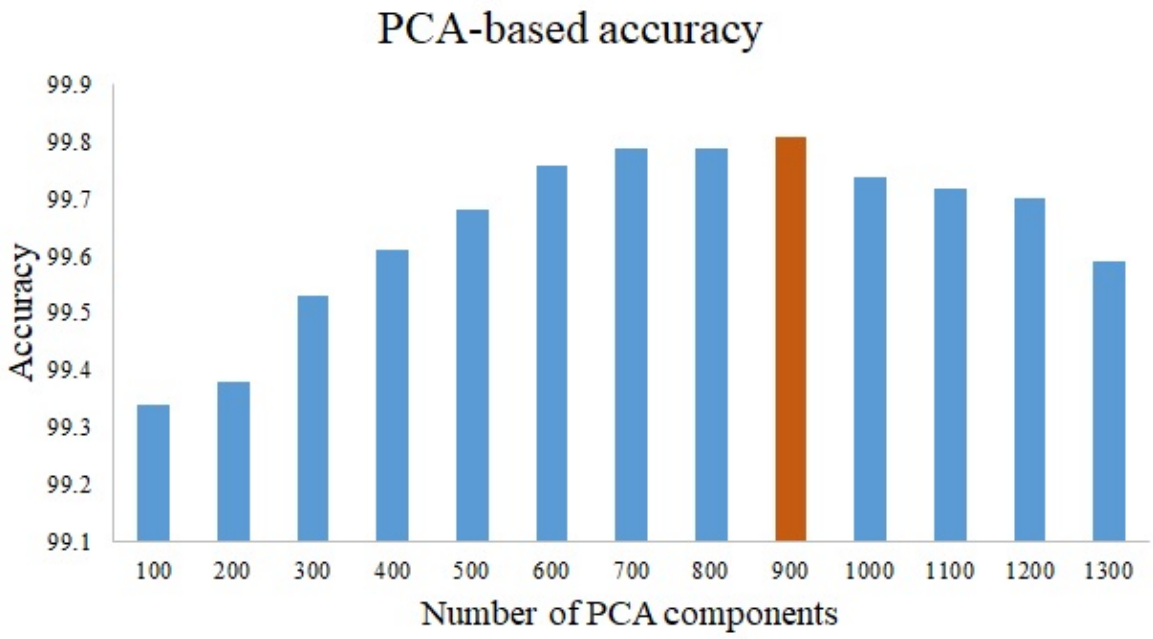
Besides, the accuracy rates have been increased by applying PCA on HOG-S vectors. This

outperforming reinforces the hypothesis of utilising the anonymised information, i.e. the utility, for HAR applications instead of using the RGB version. Thus, the original data can be dispensed, leading to a high level of privacy preservation and utility at the same time. Moreover, the proposed method proves its ability to deal with different scenarios of visual information. These scenarios include multi-view and real-world video sequences in order to pay attention to applying this method in real life. The proposed method has shown an improvement compared to the state of the art methods for five datasets by 3.04%, 3.14%, 0.83%, 3.67%, and 16.71% for DHA, KTH, UIUC1, UCF sports, and HMDB51 datasets, respectively, and a comparable accuracy rate for Weizmann dataset.

However, the accuracy rates can be improved by addressing the global motion in some of the datasets, *i.e.*, KTH, UCF sports and HMDB51, since our proposed algorithm for computing the temporal saliency (Chapter 3) avoids this issue. This global motion is included in the temporal salience map and also it contributes in calculating the feature vector which in turn affects the rate of recognition.

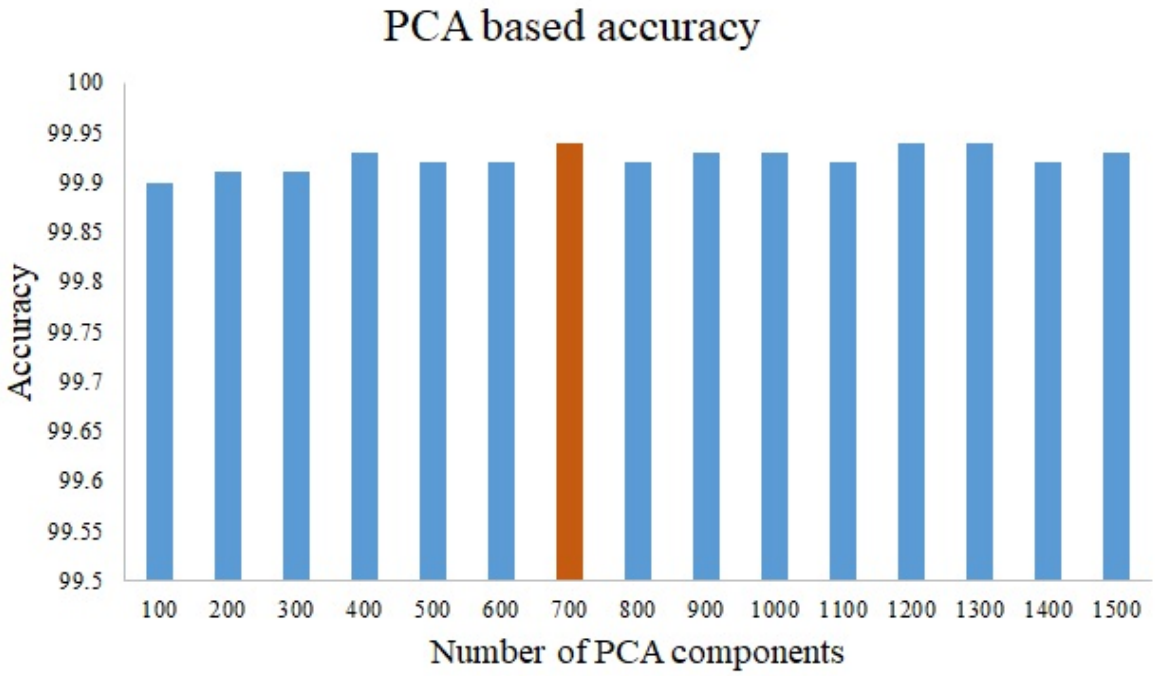


(a) Using KNN classifier

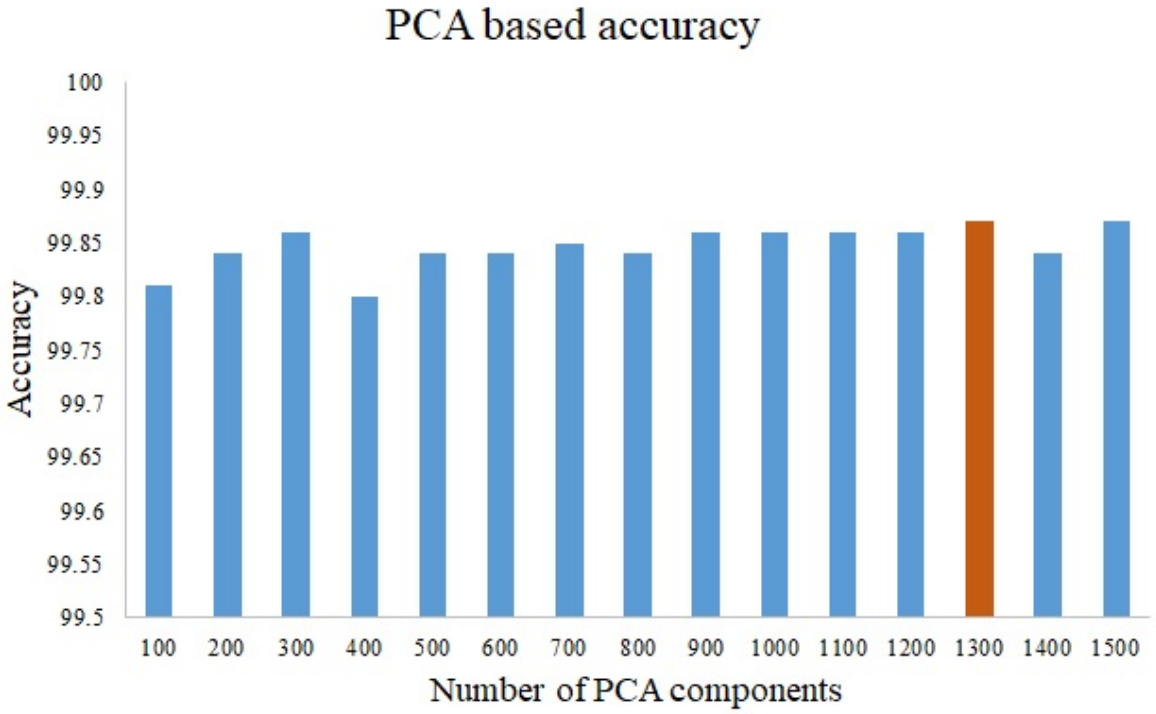


(b) Using QSVM classifier

Figure 4.14: PCA components-based accuracies of Weizmann dataset.

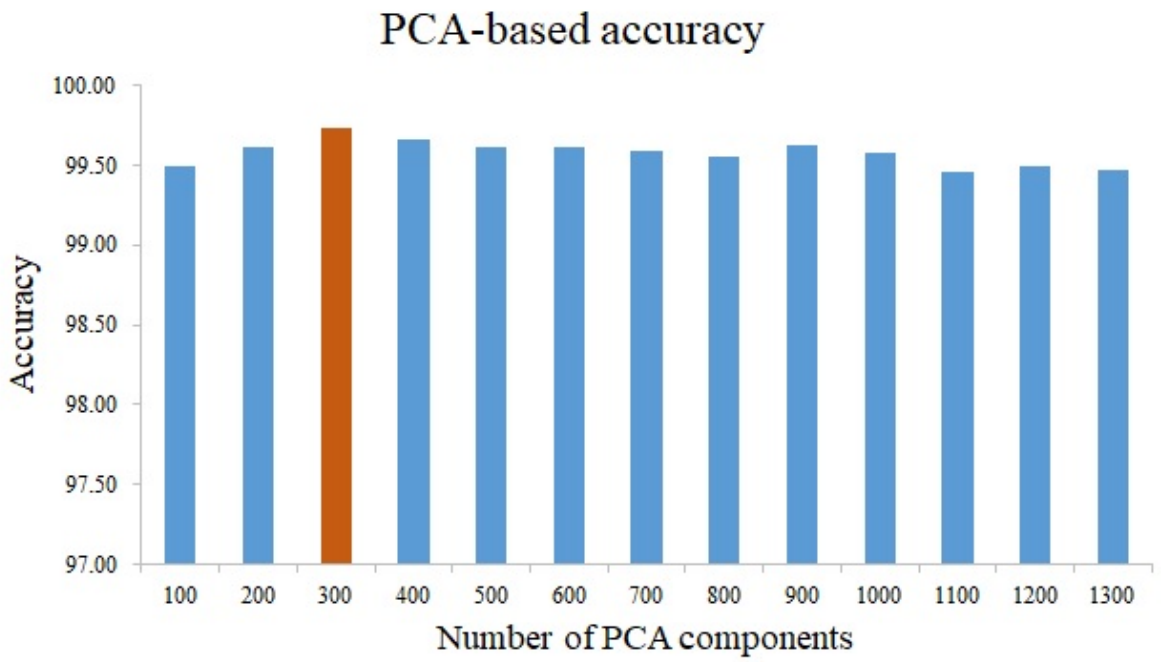


(a) Using KNN classifier

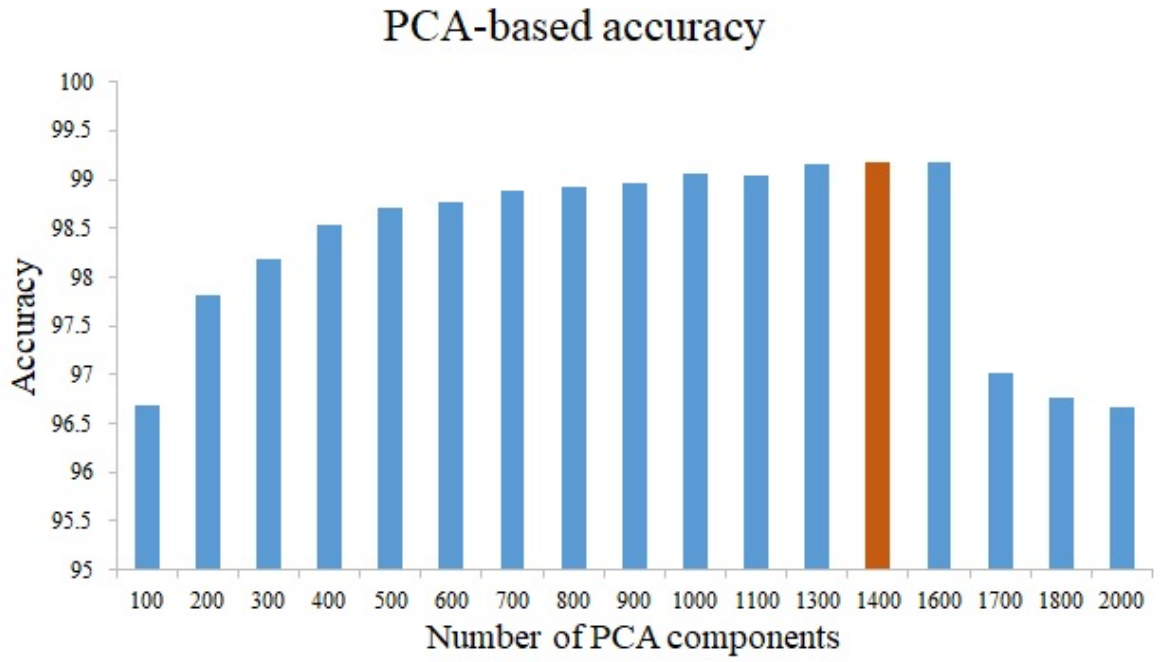


(b) Using QSVM classifier

Figure 4.15: PCA components-based accuracies of KTH dataset.

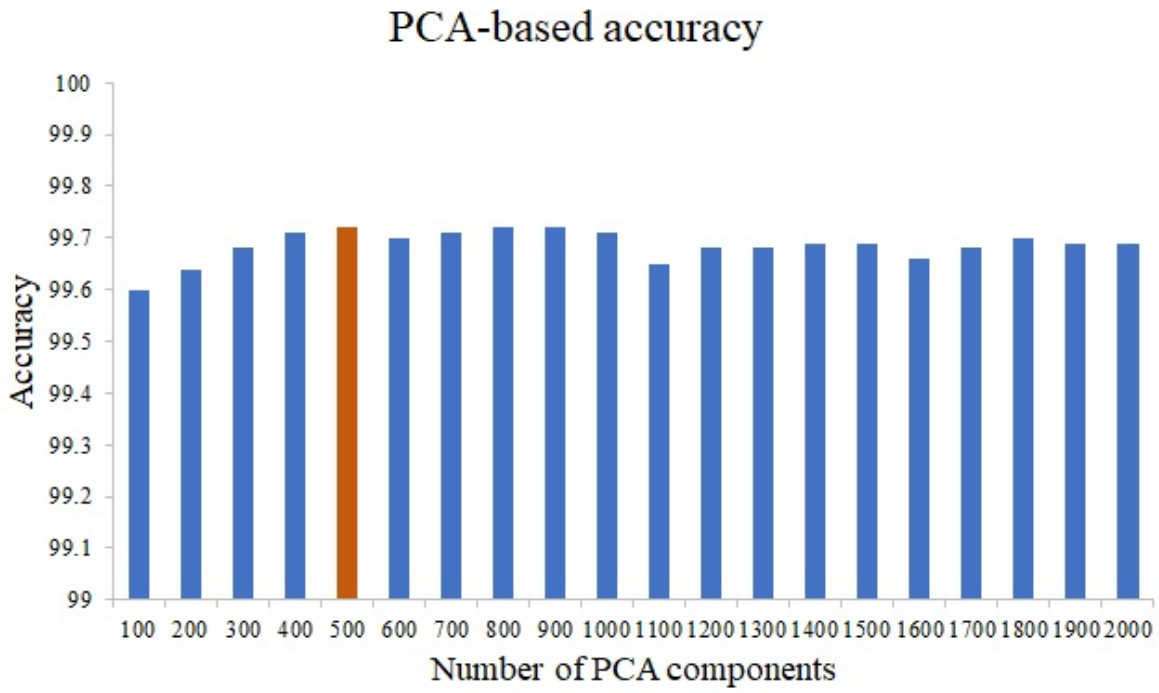


(a) Using KNN classifier

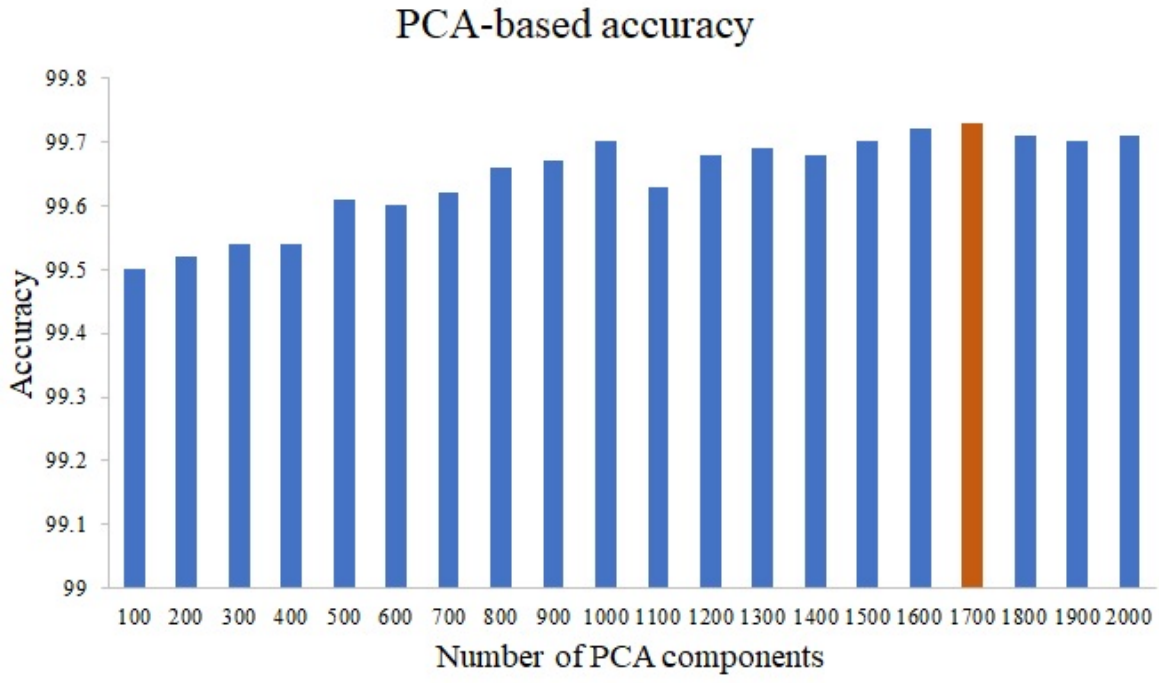


(b) Using QSVM classifier

Figure 4.16: PCA components-based accuracies of DHA dataset.

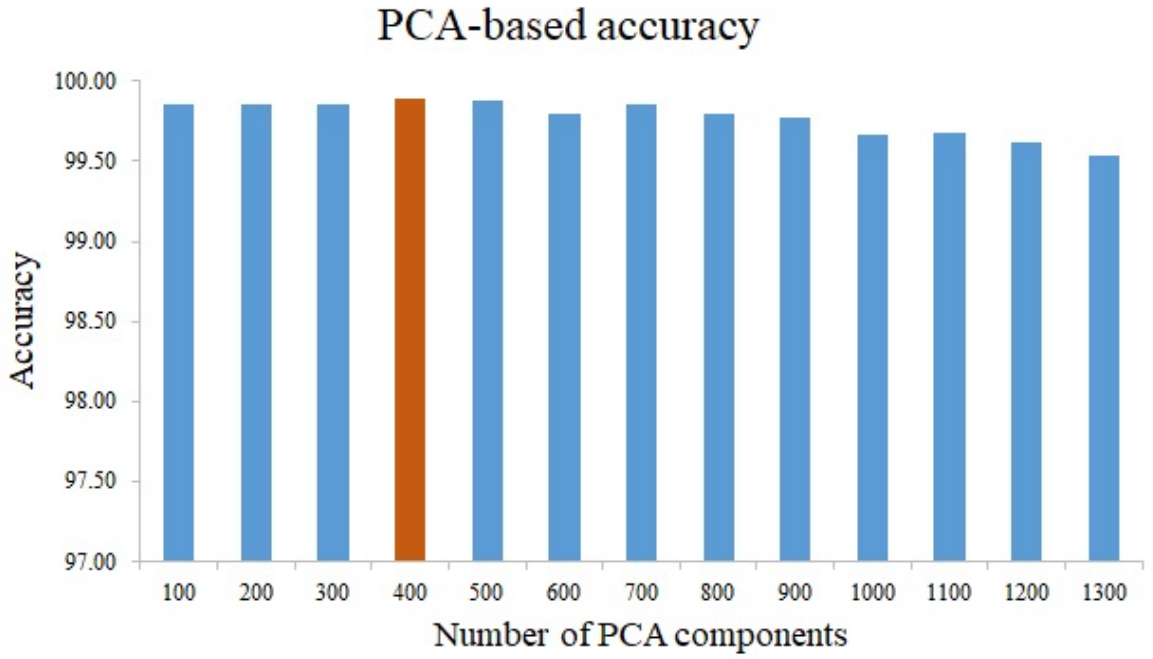


(a) Using KNN classifier

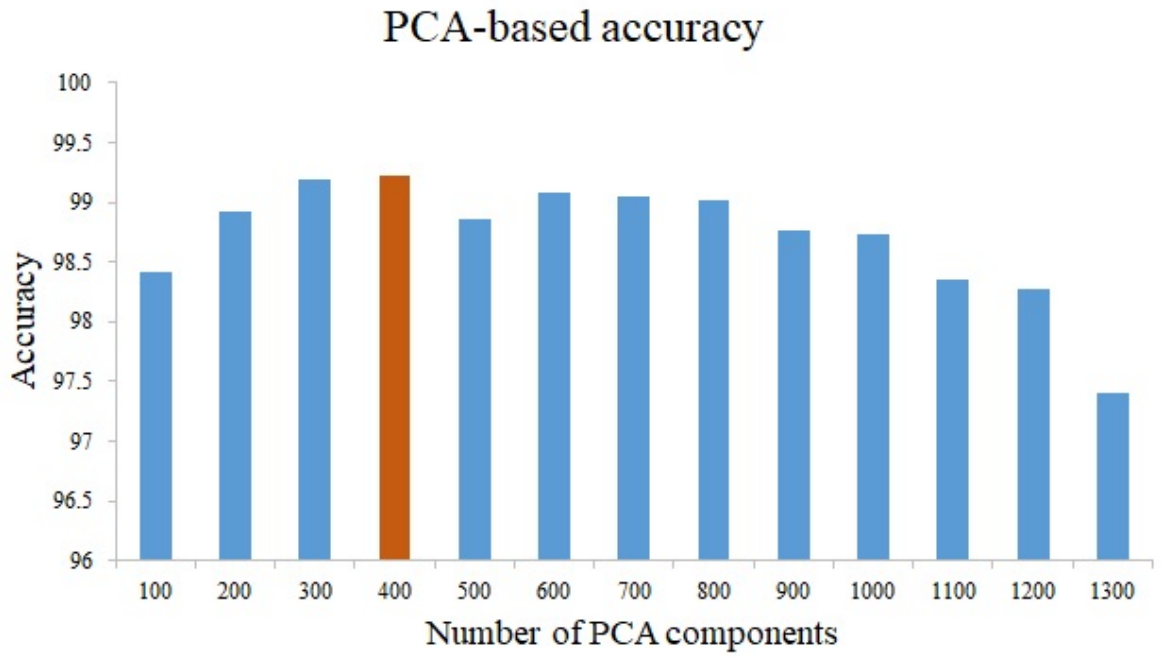


(b) Using QSVM classifier

Figure 4.17: PCA components-based accuracies of UIUC1 dataset.



(a) Using KNN classifier



(b) Using QSVM classifier

Figure 4.18: PCA components-based accuracies of UCF sports dataset.

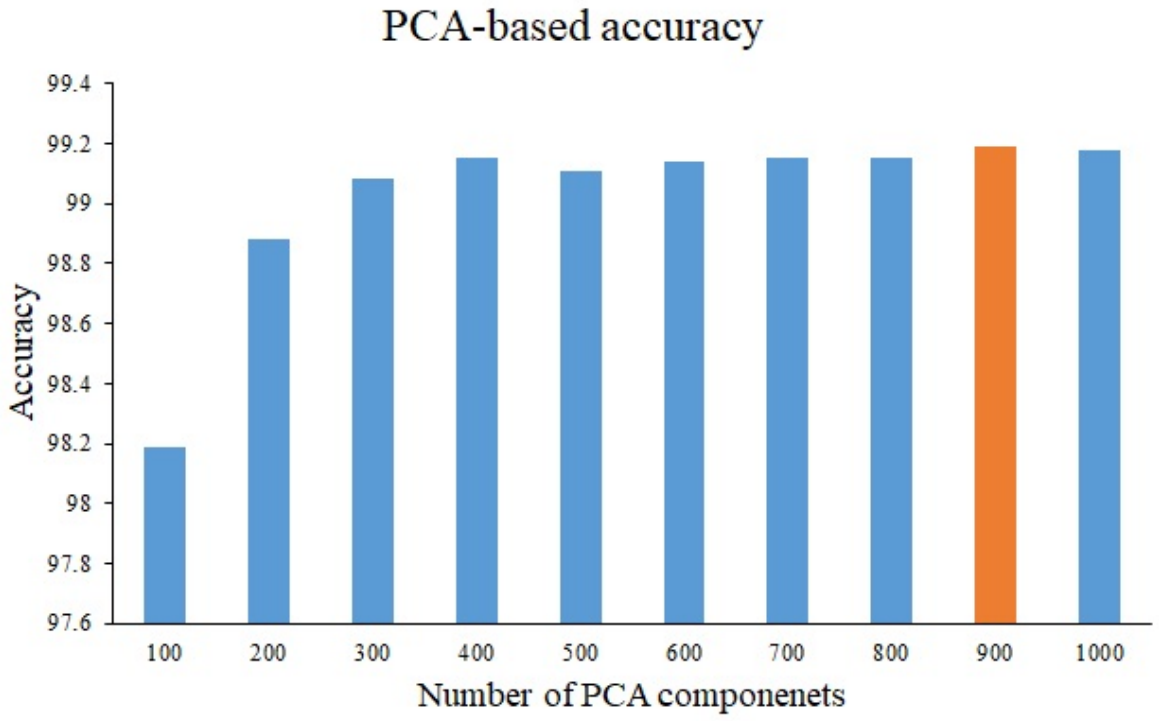


Figure 4.19: PCA components-based accuracies of HMDB51 dataset using KNN classifier.

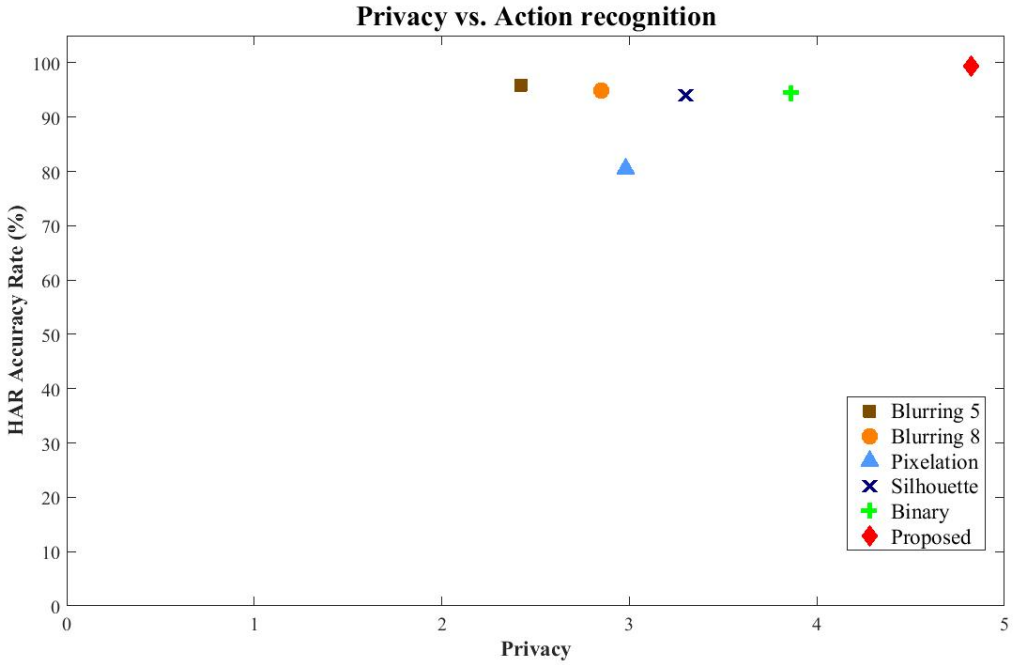


Figure 4.20: Privacy vs. action recognition based on the video-based anonymisation methods.

Chapter 5

Neuromorphic domain Human Action Recognition

5.1 Introduction

The standard active pixel sensor (APS) performs well in monitoring the human in AAL [16, 219, 220, 221], since action recognition is mainly classified using the APS by exploring the visual information in video sequences. These actions are typically represented by detecting a set of features. Although APS-based HAR has been successful, acquiring the actions by APS devices involves many limitations [20, 23]; for instance: limited frame rate, redundancy between the successive frames, motion blurring, and also the APS camera has a high power consumption. Some of these problems may be exacerbated when applied in the real world. Besides all these problems, privacy is lost in these systems, and people are concerned about using the APS for in-home monitoring [14, 11].

Considering the limitations mentioned above, a new neuromorphic vision sensing (NVS) camera, which is inspired by the retina of the human's eyes, is recently innovated [19, 20]. This NVS device measures the change of intensity at each pixel at a rendering frame rate up to 2000 fps asynchronously, *i.e.*, independently, instead of acquiring the intensities with consuming low power. The intensity change is encoded at each pixel in the form of *event* or *spike*. This encoding scheme makes the NVS camera assigns an adaptive sampling rate at each pixel based on the motion in the scene instead of using a fixed sampling rate for all pixels. The use of such

independent sampling rate addresses the motion blur caused by the high-speed objects which means that the NVS based camera is a data-driven sensor [23]. Combining these characteristics in one device may make the neuromorphic camera the preferred vision sensor for robotic and mobile applications [157, 254, 255, 146].

Since the NVS generates a stream of events without intensity values, the identity details of the human cannot be recognised. This characteristic makes the NVS a useful candidate to address the issue of privacy in the application of AAL. However, the output of the NVS is different compared to the APS camera, which means that it is difficult to apply the standard vision-based algorithms to deal with the visual data since the output represents intensity changes rather intensities. Therefore, we need a new method that can explore the NVS domain-based data for achieving the utility of this new sensor beyond the anonymisation.

Recent work on exploring the neuromorphic sensing domain for HAR has been focused on involving the NVS devices for low-level semantic tasks, such as, hand tracking. Furthermore, exploring the NVS sensor for higher-level semantic tasks, such as multi-classes action recognition and behaviour understanding, is still limited. One reason for obstructing the progress in this field is due to the higher cost of the neuromorphic sensors [152], resulting in a scarcity of annotated NVS-based training datasets [144, 256]. Another reason is that the current methods of exploring the visual data are unable to deal with the events, as above mentioned.

However, there are two scenarios to address the lack in providing NVS datasets, the first solution is proposed by displaying the standard APS sequences on the screen under controlled settings and standing the NVS sensor on the opposite side to acquire the events [256, 146]. By this method, the sensor also captures the brightness changes of the underlying electronics of the screen itself and generated noisy events which affect the accuracy of identifying the content of the stream of events. To overcome this obstacle, the second category of the solutions presents software-based tools, such as in [160] and [161], to provide a useful and cheap framework to generate annotated training datasets for the higher-level action recognition tasks. These simulators guarantee that the stream of events is generated without including noisy events since the events are generated based on the log intensity differences without displaying the video sequence on a screen [161].

Considering any new research or device in the applications of HAR imposes challenges. For

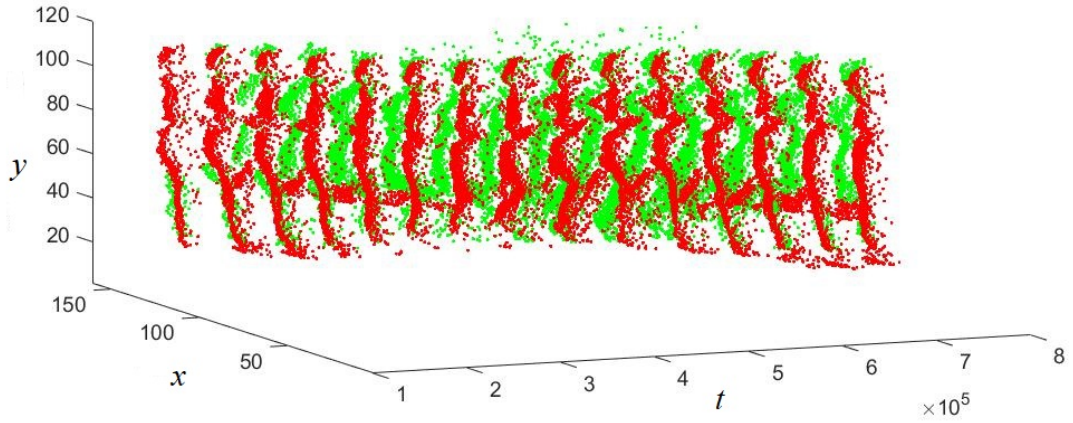


Figure 5.1: Representation of the events of a running action from KTH dataset [4] using an emulator to generate the events. Green/Red points are for visualisation of ON and OFF events.

instance, in the AAL application, privacy is one of the most critical issues that must be considered and addressed. Therefore, applying the neuromorphic sensor in AAL should consider these requirements. Since the neuromorphic sensing domain produces a stream of events instead of intensities, privacy seems to be addressed and provides a vision domain without any concern about the identity violation. Addressing the privacy issue by such sensor can be observed in Figure 5.1. This Figure shows events, which are generated in the neuromorphic domain, for a subject doing an indoor running activity from KTH dataset [4]. The events are represented in a 3D space to refer to the locations of the brightness change, spatially and temporally, without including any details about the subject's identity. These events can preserve privacy and provide an obfuscated sensing domain to be exploited in AAL. These events also have potentially valid information that can be used to recognise the action. Each event represents a log intensity change in a specific time in the scene, *i.e.*, motion. In other words, the motion is computed and included in the events and re-computing of the motion is avoided. Therefore exploring these events and their potential information in a reliable approach is an essential goal to prove the utility of the event-based domain.

In this chapter, the issue of privacy preservation is addressed by using the neuromorphic sensing domain. Mainly, this chapter emphasises on achieving the compatibility between privacy and utility. Therefore, we conduct extensive experiments on different neuromorphic scenarios in order to test the proposed method for this new domain.

The main contributions of this chapter are:

1. Exploiting the NVS domain for preserving the privacy in AAL application.
2. Proposing a new set of local and global temporal features on NVS domain for HAR.
3. Proposing a new de-noising algorithm to remove the noisy events that can be acquired by the native NVS cameras.

This chapter is organised as follows: First, Section 5.2 explains the related work on exploring the NVS domain for HAR. Second, in Section 5.3, operation of the NVS-based camera is introduced as well as the advantages and disadvantages of this sensor. Third, exploring the NVS domain data to extract meaningful descriptions to use it to recognise the actions is presented in Section 5.4. Fourth, the experimental results are shown in Section 5.5 where the proposed descriptors are tested on several standard datasets for HAR. Finally, the conclusions are drawn in Section 5.6.

5.2 Related Work

Recently, NVS has been explored for HAR with limited contributions, such as in [144, 146]. Since this sensor has never been used before in AAL applications, reviewing the current work focuses on the utility of the NVS domain for HAR.

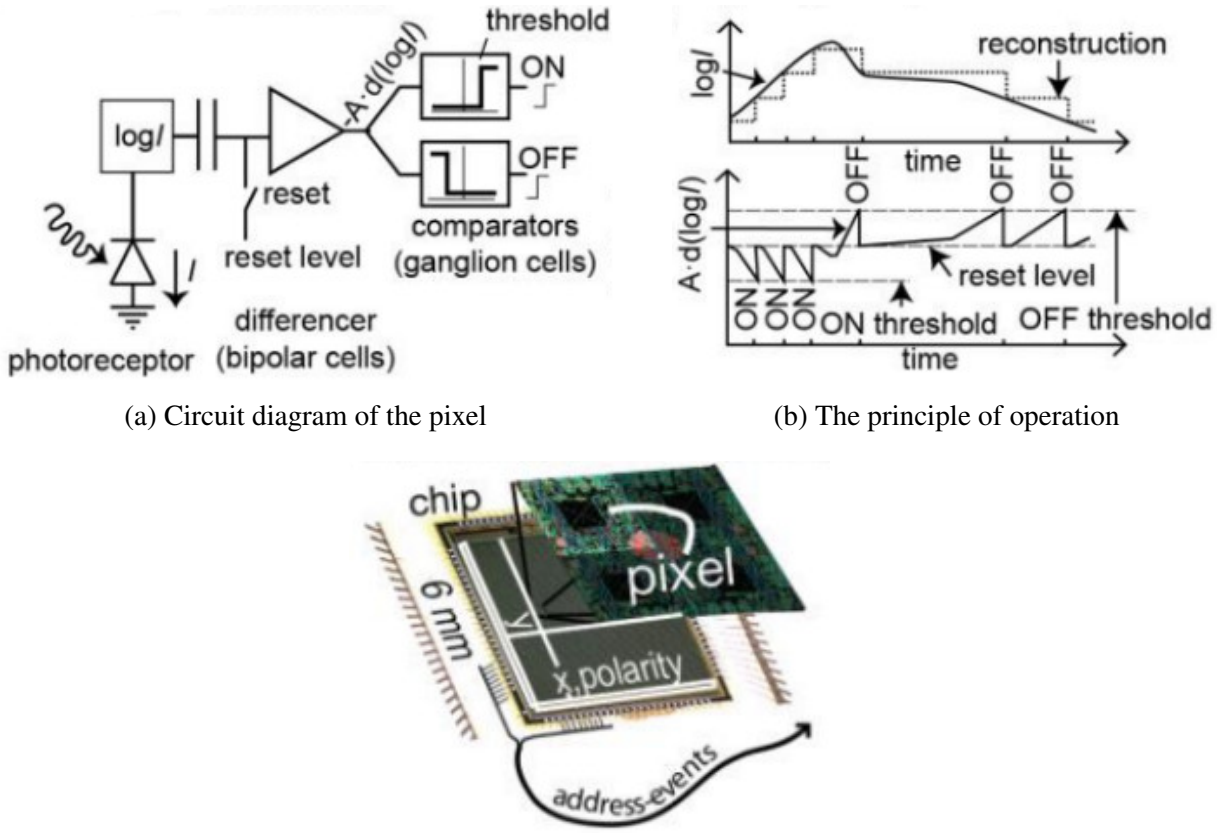
One of the problems in exploring NVS for HAR is the availability of annotated NVS training datasets [144] leading to the rarity of contributions in this area. In Chapter 2, we explained that the existing works on NVS-based HAR are categorised into two themes: behaviour monitoring [148, 147, 162, 20, 1, 164, 165, 166] and higher-level semantic action recognition [146, 144, 143, 168]. In the first category, NVS data was limited to monitor specific actions, such as finger/hand movement [147, 162, 1, 164, 165, 166], fall detection [148] and sleep monitoring [20]. The specifications inspire interest in using the neuromorphic camera in these applications that this device presents: including the low power consumption and low latency, which make this sensor suitable and functional for real-time applications that require an immediate response. The topologies of exploiting this sensor for action monitoring are a single device and stereo cameras.

In the second category, the NVS camera has been explored beyond a single action into multi-class action recognition. Specifically, the current work mostly focuses based on deep learning [146, 144, 168] or hand-crafted feature learning [143]. On the one hand, in deep learning-based methods, a set of NVS-based frames is stacked to construct a new domain-based frame and used to train the neural network. Although the existing methods, such as in [144], present good qualitative results, the accuracy rates of recognition depend on the quality of constructing the frames. On the other hand, the hand-crafted based approach constructs a set of motion maps from the NVS events to label the data and then extracts standard features to learn the classifier.

In both themes mentioned above, the events are converted into a frame-based formulation using a non-neuromorphic domain for learning, ignoring the native NVS domain for learning. These contributions are valuable efforts to develop new event-based algorithms, but these are considered computationally expensive, especially in the case of two-stream learning, such as in [144]. Furthermore, these methods focused on converting the events from the NVS domain into frame-based modelling in other domains leading to losing the nature of the events. Therefore, focusing on the NVS domain, *i.e.*, exploring the events directly, for learning, can be useful to improve the performance of recognition, and this chapter aims to verify the usefulness of the neuromorphic domain for HAR. Accordingly, we propose a new method to explore the NVS domain by considering the temporal patterns of ON and OFF events locally and globally to extract a reliable feature learning for HAR. The proposed method analyses the patterns of the polarities using only NVS domain-based events and avoids converting the events into other domains without losing the essence of neuromorphic computing.

5.3 The operation of the NVS camera

The operation of the NVS camera is illustrated in Figure 5.2. Contrary to the standard pixel-domain based camera, where the camera records the information of the pixels at a constant frame rate based on the intensities, the NVS camera acquires the luminance change instead of intensity with a variable sampling rate at each pixel. Accordingly, the event is triggered if the luminance change, *i.e.*, log intensity, exceeds the predefined threshold. Generating the events is asynchronously and independently at each pixel in the chip's array, Figure 5.2(c), where each



(a) Circuit diagram of the pixel

(b) The principle of operation

(c) Four pixels layout chip

Figure 5.2: Dynamic vision sensor: structure and operation [5].

pixel continues monitoring the brightness changes over the time and launches the event if the magnitude of the intensity change is sufficient to generate an event.

The NVS camera, also called a dynamic vision sensor (DVS), generates the events only in the case that the brightness change at a specific pixel is sufficient to exceed the predefined threshold and the pixel is set to idle if there is no luminance change. This mechanism makes the DVS camera consumes low power and avoids acquiring redundant information. In general, when an event, e, is detected, it is generated and recorded with the x and y coordinates of the corresponding pixel's location in the array and the timestamp, t, as well as the orientation of the shifted log intensity, *i.e.*, polarity p. This is the only required information to represent the e-th event. The amount of the luminance change makes the NVS sensor a data-driven device. Moreover, the NVS sensor can capture the high-speed objects by generating more events, *i.e.*, increasing the sampling rate, contrary to the slow-motion objects where the sampling rate is reduced. Therefore, the sampling rate is temporally adapted based on the density of the motion

in the scene, meaning that the rate of each pixel is independent of other pixels in the array.

The DVS camera timestamps the events in a microseconds resolution and the delay of transmission is measured in milliseconds. Finally, the generated event is then transmitted over a shared bus; namely, the address-event bus (AER) [257, 258].

5.3.1 Modelling the event

As mentioned in the introduction, the pixels in the NVS camera respond independently to the change in the brightness [150], such as the log intensity, $\mathcal{L}$ of pixel P_k is

$$\mathcal{L}(P_k) = \log(I_k), \tag{5.1}$$

where I_k is the intensity of pixel P_k. Accordingly, an event $e_k = (x_k, y_k, t_k, p_k)$ is generated at pixel P_k with coordinates (x_k, y_k) at time t_k when the magnitude of $\mathcal{L}(P_k)$ is shifted since the last event at P_k, i.e.

$$\Delta\mathcal{L}\left(x_k, y_k, t_k\right) = \mathcal{L}\left(x_k, y_k, t_k\right) - \mathcal{L}\left(x_k, y_k, t_k - \Delta t\right), \tag{5.2}$$

overcomes a temporal contrast threshold, $\pm\Theta$, where $\Theta > 0$. Δt is the time when the pixel P_k is idle since the last event at P_k. When the log intensity at P_k exceeds Θ, e_k is triggered with the orientation of log intensity change, i.e. polarity $p_k \in \{-1, 1\}$.

Refer to Eq. (5.2); we notice that this formula is similar to find the pixel difference between successive frames. This means that the log intensity is evidence of including the motion in the events. Therefore, the motion can be instantly represented instead of computing this motion again [259], reducing the complexity of exploring the events.

5.3.2 The advantages of the NVS camera

The NVS sensor offers several advantages over the standard vision sensor:

- *Temporal Resolution*: The sensor presents a high temporal resolution in microseconds compared to the conventional camera, which has a temporal resolution measured in seconds. This temporal resolution makes the NVS camera captures fast changes in the bright-

ness of high-speed objects without blurring noise which represents one of the challenges in the frame-based vision cameras.

- *Low Latency*: The change in the intensity is captured and the event is generated instantaneously with delay in microseconds (about 10 μs [23]). Thus, the events are generated asynchronously without wasting time waiting for other events to trigger.

- *Low Power*: The NVS sensor produces and processes only the events of the brightness changes without redundancy, leading to consuming low power. The power consumption in some NVS cameras can be measured in μW [260, 1], which is much less than the power consumed by the standard cameras.

- *High Dynamic Range*: The sensor has a high dynamic range up to more than 120 dB versus 60 dB of the standard cameras since the events can be generated regardless of the illumination conditions.

These advantages make the neuromorphic sensor useful for mobile applications, such as robots, and it is also functional in the environments where the light condition is uncontrolled and cluttered.

5.3.3 The disadvantage of the NVS camera

Though attractive characteristics of the NVS cameras, a number of challenges are included in this sensor:

- *Cost*: Commercially, the neuromorphic sensors are costly compared to the standard cameras [152]. For instance, the price for the DAVIS346 camera is around six thousand dollars [159] following the low production of this type of sensors. Therefore, only a few research groups can afford this sensor.

- *Resolution*: Because the vision chip of the sensor array of pixel photo-circuits and this pixels' array occupies a large size in the hardware, the resolution of this sensor is restricted to limited dimensions. For instance, the recent DVS-Gen2 has a 640×480 resolution [153], which is considered the highest resolution so far.

- *Noise*: Like any vision sensor, the noise can be generated following the noise of the transistor circuit, leading to generating noisy events which requires designing algorithms to eliminate there events.

Others, such as the complexity of design and reliability of processing, with the limitations mentioned above, restrict the usage and benefit of such sensor.

5.4 The proposed method

In this section, we present the proposed method to build two descriptors to explore NVS domain-based data. Figure 5.3 depicts the pipeline of the proposed method exploring the NVS domain for HAR. In the following, we explain the steps of processing the events to extract features in order to test the utility of the NVS domain for HAR.

5.4.1 Pre-processing the noisy events

Depending on the threshold magnitude, some events are recorded in isolation without leading to any semantic meaning. These events can be caused by various reasons, such as, the local

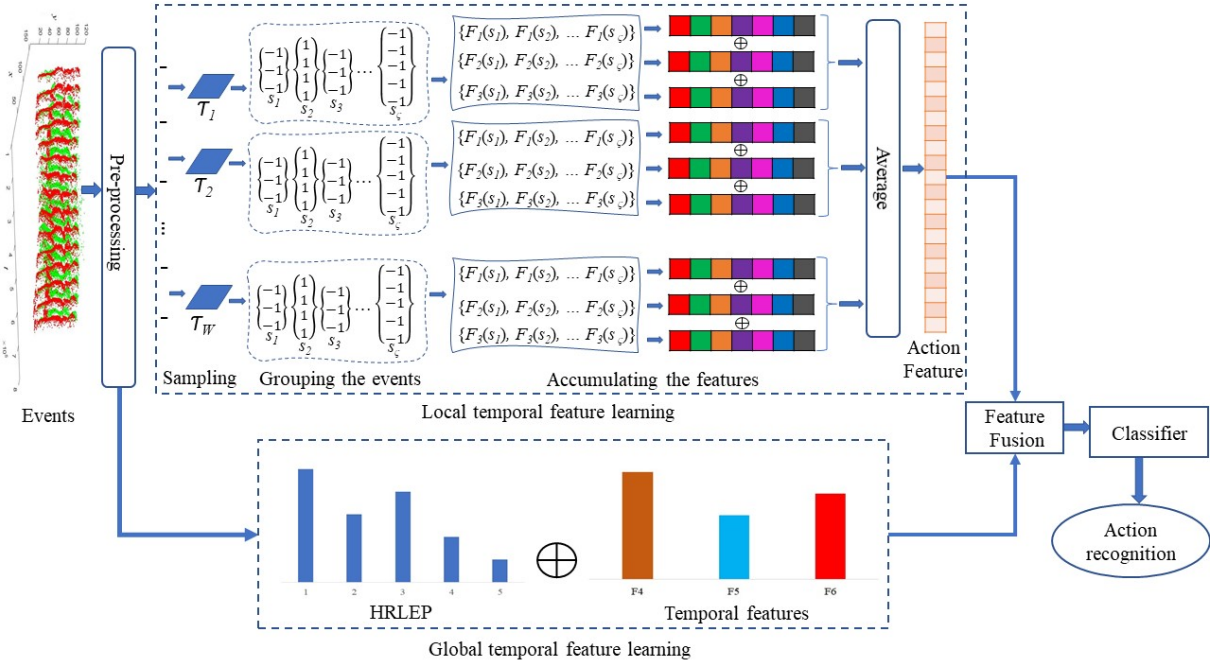


Figure 5.3: The pipeline of the proposed method to explore the NVS domain for HAR.

light change in the scene or the flickers of the electronic parts. We denote such events as noisy events and a pre-processing step for removing such events (de-noising) is applied on the events stream. In the proposed method, de-noising is applied on the events that are generated by using the native DVS camera because the native neuromorphic camera generates a lot of such noisy events compared to other scenarios.

Let $\mathbb{E} = \{e_n | e_n = (x_k, y_k, t_k, p_k)$, and $1 \leq n \leq N$, is a stream of events, where N is the length of the event stream. $\mathbb{E}$ is partitioned into time slices, $\mathbb{T} = \{\mathcal{T}_w\}_{w=1}^{W}$, where $\mathcal{T}_w$ is the time slice w. This partitioning is based on the principle of the frame rate that one would expect for a conventional camera video sequence. For example, if we have an NVS stream for 5 seconds, we generate 150 event slices assuming a 30 frames per second frame rate.

After partitioning the stream into event slices, for each slice let $\mathbf{E}_w = \{e_\ell | e_\ell = (x_\ell, y_\ell, t_\ell, p_\ell)\}$, and $1 \leq \ell \leq L$, be the event stream in slice w, where $\mathbb{L}$ is the length of the event stream in a slice, the following operations are applied. For each event e_ℓ at spatio-temporal location (x_ℓ, y_ℓ, t_ℓ), a 3×3 window on xy plane centred on the event location (x_ℓ, y_ℓ, t_ℓ) is considered and the number of events $C_{\ell_{(x,y)}}$ recorded on each of nine spatial coordinates (x, y) of the window over the total time of the slice is counted. This is followed by computing the total number of events in the 3D window-slice, S_ℓ, and the maximum events over the slice length, m_ℓ, as follows:

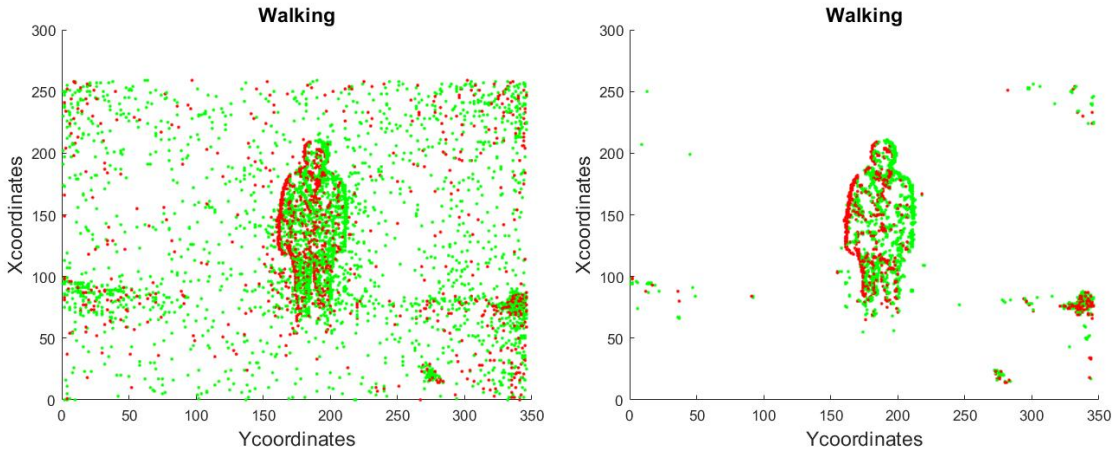
$$S_\ell = \sum_{i=x_\ell-1}^{x_\ell+1} \sum_{j=y_\ell-1}^{y_\ell+1} C_{\ell_{(x,y)}}. \tag{5.3}$$

$$m_\ell = \max_{i=x_\ell-1, j=y_\ell-1}^{i=x_\ell+1, j=y_\ell+1} C_{\ell_{(x,y)}}. \tag{5.4}$$

Finally, e_ℓ is processed to obtain new polarity, $\acute{p}_\ell$, of the event as follows:

$$\acute{p}_\ell = \begin{cases} p_\ell & \text{if} \quad S_\ell \leq (k \times 3 \times 3 \times m_\ell), \\ 0 & otherwise, \end{cases} \tag{5.5}$$

where $\{k \in \mathbb{R}^+ | k \leq 1\}$ is a user defined parameter for controlling the number of events to be removed. Figure 5.4 shows an example for de-noising a sampled slice from walking action recorded by a native neuromorphic camera from the dataset in [6] using $k = 0.25$. In this case the pre-processing step has resulted in removing approximately 70% of events in each slice and



(a) Before de-noising (b) After de-noising

Figure 5.4: Event de-noising based on adaptive masking applied on a stream of events from the walking action. The time slice for the events in this figure is 0.03 second and $k = 0.25$.

retaining the relevant events of the action.

We observe that the native camera generates a large number of noisy events. These noisy events can affect the exploitation of the exploiting the NVS domain for feature extraction. The effect of these noisy events on action recognition task will be tested in the evaluation section. The de-noising is applied on the native DVS based dataset since the emulator-based datasets output a neuromorphic domain without such noisy events.

5.4.2 Local temporal feature extraction

An action event stream can be represented considering the overall spatio-temporal patterns that appear in the overall action sequence, as well as considering the local variations corresponding to the actions. In this section we address how to extract local features from the events stream, considering the events in partitioned time slices, $\mathcal{T}_w$. Since each action results in different spatio-temporal patterns of events at each time window, the local descriptors aim to recognise these patterns leading to representing discriminating features for specific action streams.

The process is started with $\mathbf{E}_w$ at $\mathcal{T}_w$ by sorting all e_ℓ in the ascending order of the x coordinate followed by grouping these events in $\mathcal{T}_w$ into $\{s_g|1 \leq g \leq G\}$, where s_g defines ρ events that are successive and have the same polarity, such that:

$$s_g = \{e_i | e_i = (x_i, y_i, t_i, p_i), and\ 1 \leq i \leq \rho\}, \tag{5.6}$$

where $x_{i+1} \geq x_i$ and $p_{i+1} = p_i\ \forall i$. According to Eq. (5.6), all events in s_g represent a pattern of log intensity change. Processing such patterns of polarities contributes to tracking the dynamic changes for each action and capturing the local structure of the events. This is achieved by modelling these changes in terms of the relationship of horizontal and vertical locations, *i.e.*, (x, y) coordinates of the events in each set, s_g in terms of the following quantities:

$$m_g = \mu_x(s_g) - \mu_y(s_g), \tag{5.7}$$

$$v_g = \sigma_x^2(s_g) - \sigma_y^2(s_g), \tag{5.8}$$

$$d_g = \sigma_x(s_g) - \sigma_y(s_g), \tag{5.9}$$

where μ, σ^2 and σ are the mean, variance and the standard deviation of the spatial coordinate x and y of the events in s_g, respectively. This gives us three data vectors, $\mathbf{M}_w = \{m_g | 1 \leq g \leq G\}$, $\mathbf{V}_w = \{v_g | 1 \leq g \leq G\}$ and $\mathbf{D}_w = \{d_g | 1 \leq g \leq G\}$, for each $\mathcal{T}_w$. Then these data vectors are transformed into three vectors containing higher order statistics of the data vectors as follows:

$$F_{1_w} = \left[\mu(\mathbf{M}_w), \max(\mathbf{M}_w), \min(\mathbf{M}_w), \sigma(\mathbf{M}_w), \sigma^2(\mathbf{M}_w), \gamma(\mathbf{M}_w), \kappa(\mathbf{M}_w)\right], \tag{5.10}$$

$$F_{2_w} = \left[\mu(\mathbf{V}_w), \max(\mathbf{V}_w), \min(\mathbf{V}_w), \sigma(\mathbf{V}_w), \sigma^2(\mathbf{V}_w), \gamma(\mathbf{V}_w), \kappa(\mathbf{V}_w)\right], \tag{5.11}$$

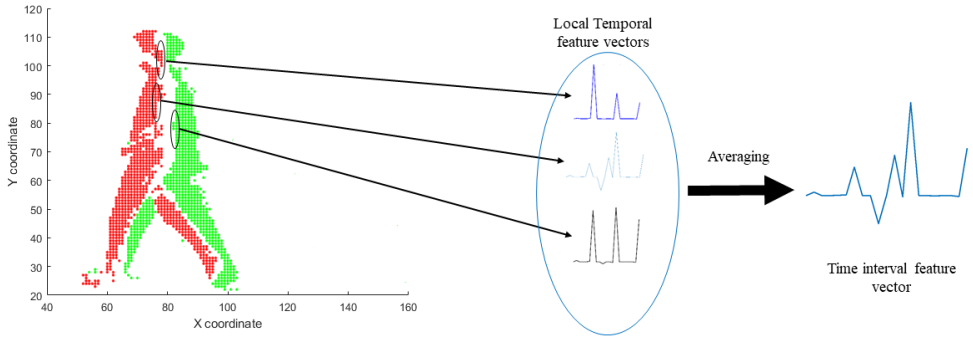
$$F_{3_w} = \left[\mu(\mathbf{D}_w), \max(\mathbf{D}_w), \min(\mathbf{D}_w), \sigma(\mathbf{D}_w), \sigma^2(\mathbf{D}_w), \gamma(\mathbf{D}_w), \kappa(\mathbf{D}_w)\right], \tag{5.12}$$

where γ and κ denote the skewness and the kurtosis, respectively. Then, for each element in feature vectors, F_{1_w}, F_{2_w} and F_{3_w}, the average over all W slices are computed to get the average feature vectors, F_{1_w}, F_{2_w} and F_{3_w}, respectively. An example on extracting these three vectors for

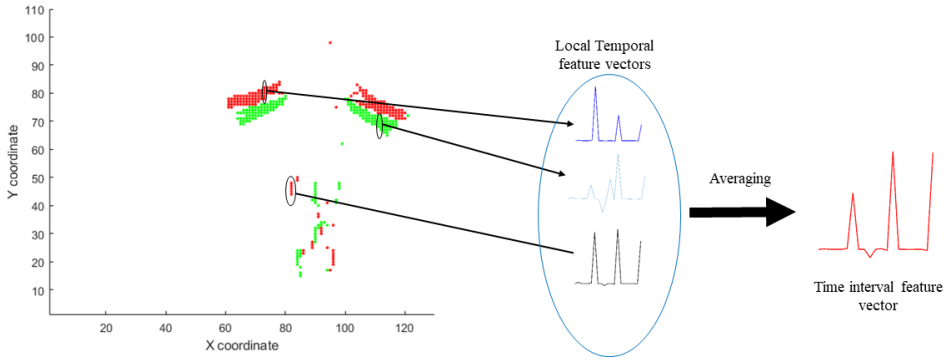
two different actions from E-KTH dataset is shown in Figure 5.5. This figure displays examples of time interval-based events which are corresponding to a single RGB frame and converting these events into local temporal features using the spatial coordinates of the polarities. These three feature vectors are concatenated to get the local feature vector, $\mathcal{F}_S = \{F_{1_w}, F_{2_w}, F_{3_w}\}$, with 21 feature elements for the event stream $\mathbb{E}$. As an example, these feature vector elements for six sequences of one of the datasets (E-KTH) in Figure 5.6.

5.4.3 Global temporal feature extraction

The second descriptor has been proposed to collect the discriminative features over the whole event stream without resorting the event into time-based slices. Global features are extracted by considering the event stream for an action as a whole without resorting it into time-based slices. On the spatio-temporal event space, for each spatial coordinate (x, y), all temporal events are



(a) Walking



(a) Waving

Figure 5.5: Two examples of extracting the local temporal feature vectors by applying the aforementioned procedure. Green and red dots represent ON and OFF polarities, respectively, and they are used here for the visualisation.

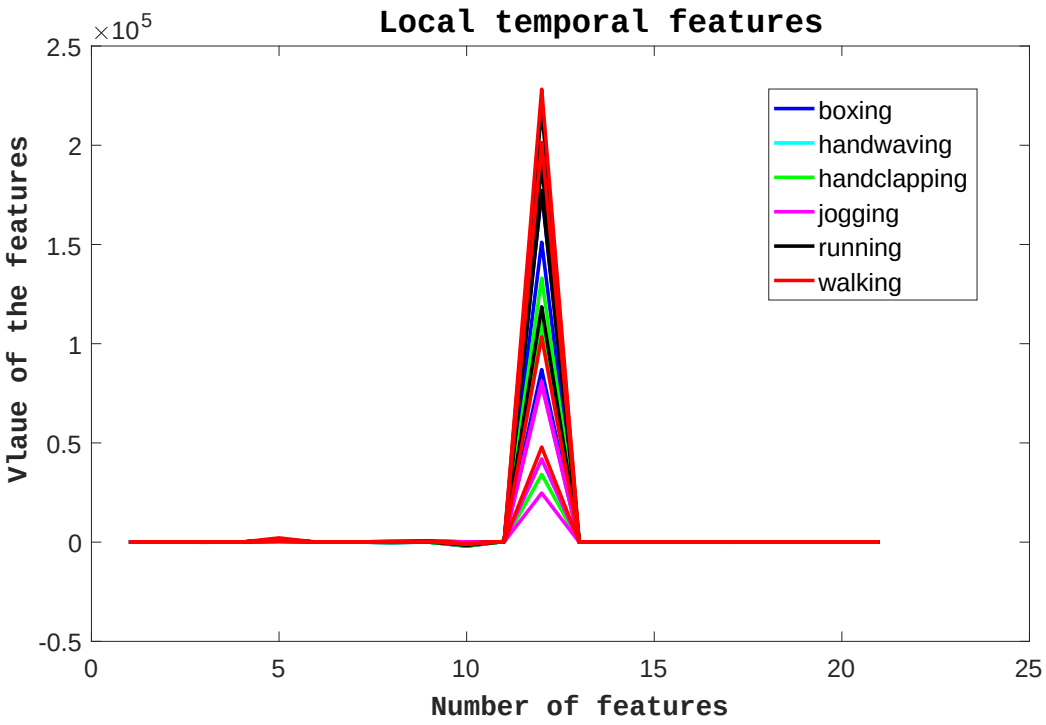


Figure 5.6: Local features for six human actions in KTH dataset: boxing, hand waving, hand clapping, jogging, running and walking. Each action is represented by four sequences.

stacked into temporal groups, $\mathcal{H}_{\mathbb{E}} = \{\delta_h | 1 \leq h \leq H\}$, where H is the total number of temporal groups for the given (x, y). A group is defined as the continuous occurrence of events (either $p_l = +1$ or $p_l = -1$) at user-specified temporal sampling periods. The minimum events for a group are considered as 2, while just the isolated single events are disregarded as noise. For all events in δ_h, the consecutive similar polarity counts are recorded as run-length encoding (RLE). RLE keeps only the counts of consecutive occurrences without the keeping the magnitudes of the polarities. Run lengths of all $\mathcal{H}_{\mathbb{E}}$ for all spatial locations are collected as a set, $\mathbb{R}$.

The first part of the global feature vector represents $\mathbb{R}$ by computing the histogram of run-length encoded polarities (HRLEP), $\mathbb{H}$. Our experiments have found that partitioning $\mathbb{H}$ into 5 bins is sufficient to capture the discriminative features from $\mathbb{R}$. In addition, we use the following global statistics considering both $\mathbb{R}$ and $\mathbb{E}$.

1. The maximum value obtained by RLE, as follows:

$$F_4 = max(R). \tag{5.13}$$

2. The maximum timestamp in seconds

$$F_5 = max(W). \tag{5.14}$$

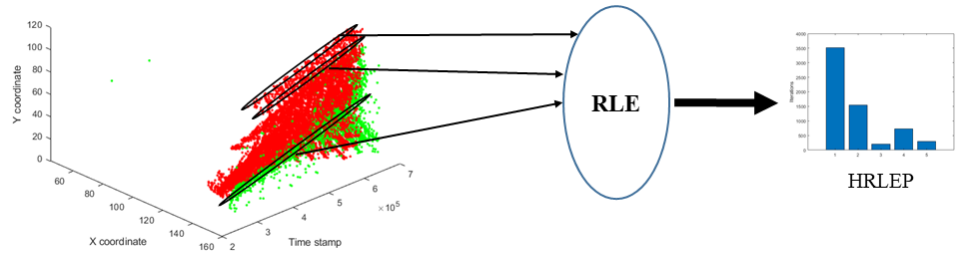
3. The number of ON events in $\mathbb{E}$

$$F_6 = \sum_{l=1}^{|\mathbb{E}|} p_l, \text{if } p_l \equiv +1. \tag{5.15}$$

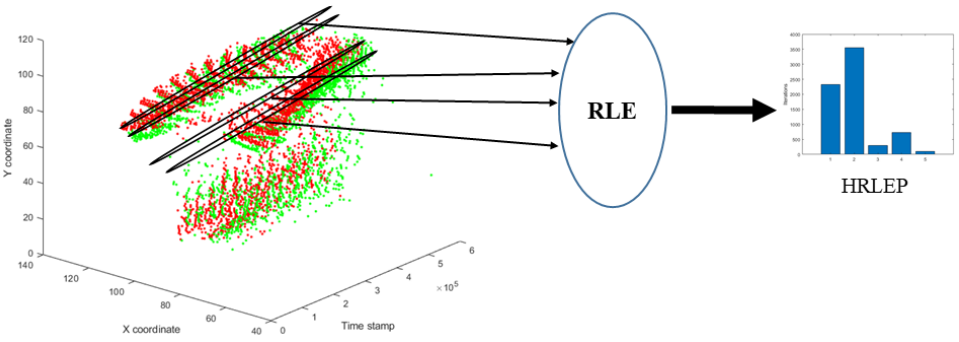
4. The number of OFF events in $\mathbb{E}$

$$F_7 = \sum_{l=1}^{|\mathbb{E}|} |p_l|, \text{if } p_l \equiv -1. \tag{5.16}$$

These four feature element with a five bin $\mathbb{H}$ produce a nine dimension global temporal feature vector, $\mathcal{F}_T(\mathbb{E}) = \{\mathbb{H}, F_4, F_5, F_6, F_7\}$, for the whole events stream. Figure 5.7 shows two examples of tracking the events at each pixel for two different actions and extracts the



(a) Walking



(a) Waving

Figure 5.7: Examples of building HRLEP features for two different actions.

HRLEP for these actions.

The global features that are extracted from four samples of action sequences of the neuromorphic version of KTH dataset are shown in Figure 5.8. The first five features in Figure 5.8 represent $\mathbb{H}$, and the other features, *i.e.*, F_4, F_5, F_6, and F_7, are represented in the indexes 6 to 9 in Figure 5.8. On the one hand, we observe that the HRLEP descriptor achieves higher discrimination for most actions compared to other components in $\mathcal{F}_T(\mathbb{E})$. Furthermore, we notice that $\mathbb{H}$ has weak discrimination for the jogging and the running actions since these two actions have a similarity in behaviour. On the other hand, the features F_4, F_5, F_6, and F_7 have a lesser level of discrimination, however, these features are concatenated with $\mathbb{H}$ to obtain a more reliable feature vector.

These feature vector elements have different dynamic ranges and some of them have values near 0 value, *i.e.*, 4, 5, 6 and 7, in Figure 5.8. Therefore, these four elements are plotted in a separate figure in order to show them clearly. Figure 5.9 shows the feature vector elements with

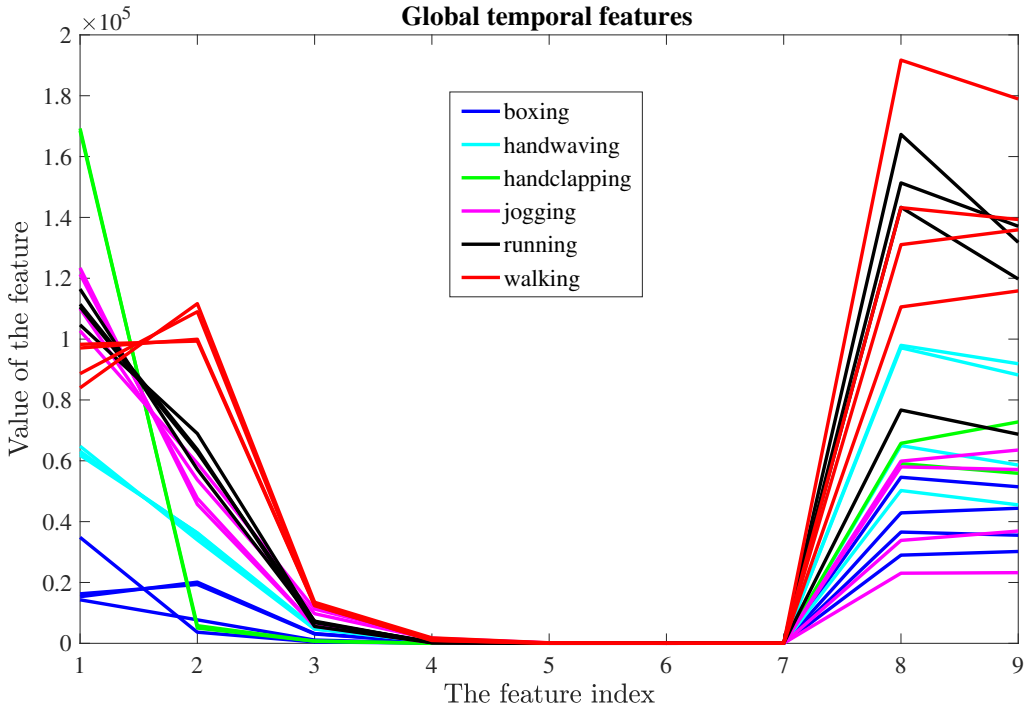


Figure 5.8: Global features for different human actions in KTH dataset for 6 actions: boxing, hand waving, hand clapping, jogging, running and walking. Each action has been represented by four sequences. Features 4, 5, 6 and 7 have a little value of magnitudes near 0, therefore, these features seem to be null in this figure.

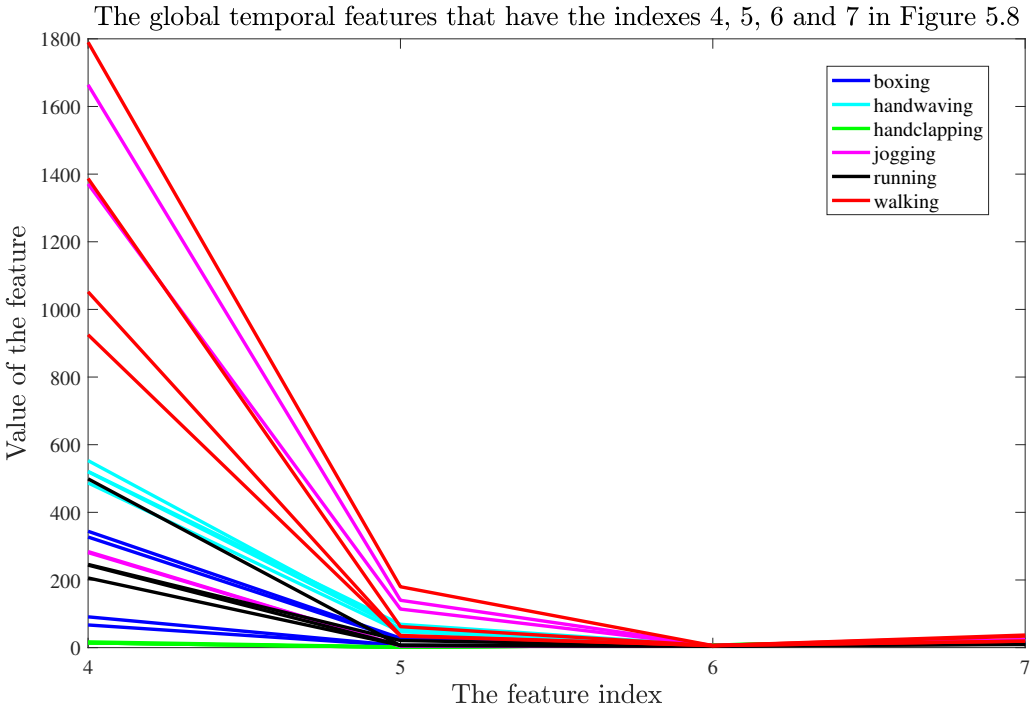


Figure 5.9: The global temporal features that have the indexes 4, 5, 6 and 7 in Figure 5.8. These features are represented.

indexes 4, 5, 6 and 7 in Figure 5.8. We observed that these elements can discriminate among the actions and there are several examples for the discrimination in Figure 5.9, such as the discrimination between handclapping and handwaving actions and the discrimination between running and walking actions.

5.4.4 Feature fusion

Finally, both $\mathcal{F}_S(\mathbb{E})$ and $\mathcal{F}_T(\mathbb{E})$ are fused to construct a local and global temporal feature vector, $\mathbb{F}(\mathbb{E})$, that overcomes the shortcomings of each descriptor, such as:

$$\mathbb{F}(\mathbb{E}) = \{\mathcal{F}_T(\mathbb{E}), \mathcal{F}_S(\mathbb{E})\}. \tag{5.17}$$

This $\mathbb{F}(\mathbb{E})$ is a 30 dimensions feature vector to represent the action in $\mathbb{E}$, and it is used to train the classifier to improve the accuracy. This feature learning method aims to emphasize that exploring the NVS domain leads to a higher level of usefulness, *i.e.*, utility, of this new vision domain.

5.4.5 Classification

For the multi-class classification problem, two classifiers were tested, which are KNN and QSVM. Each classifier will be used to classify three feature vectors, $\mathcal{F}_S(\mathbb{E})$, $\mathcal{F}_T(\mathbb{E})$, and $\mathbb{F}(\mathbb{E})$,separately, to find the best formulated features.

5.5 Performance evaluation

In this section, several experiments have been conducted using challenging datasets in the computer vision to demonstrate the utility of NVS domain. These experiments show the level of utility that is achieved using this new domain of anonymity.

5.5.1 Datasets and experiments set up

To explore the proposed method, we partitioned the experiments into three scenarios based on the method of obtaining the neuromorphic data. These scenarios are:

1. Emulator-based: In this scenario, the neuromorphic data is generated by using an emulator. There are several emulators, such as, PIX2NVS [161], pyDVS [160] and ESIM [159], that are designed to simulate the native DVS cameras. In our experiments, PIX2NVS emulator is used to generate the events from the video sequences since the only published results on using the emulators are based on PIX2VNS. For this purpose, we used four datasets; which are KTH, UCF11, UCF50 and HMDB51 and converted them into the neuromorphic version. We call these datasets E-KTH, E-UCF11, E-UCF50 and E-HMDB51 during the experiments.

2. Recording-based: We used the available dataset which is recorded from the UCF50 dataset [256] using DAVIS346redColor neuromorphic camera, and we call it R-UCF50 in this chapter. This dataset is a real-world scene containing 10 human actions: arm-crossing, getting-up, kicking, picking-up, jumping, sitting-down, throwing, turning around, walking, and waving. The details of this dataset are explained in Table 5.1.

Table 5.1: Characteristics of the two neuromorphic datasets acquired by neuromorphic devices in two different scenarios.

Name	No. of sequences	Scenarios	Resolution	No. of classes
N-Actions [6]	450	Office	346 × 260	10
R-UCF50 [256]	6681	YouTube	240×180	50

3. Native DVS-based: The DVS camera is used to acquire a real NVS dataset, which is published in [6]. This dataset is obtained by recording 10 real human actions in an office environment using DVS240C camera. We call this dataset N-Actions during the experiments. The details of this dataset are also illustrated in Table 5.1.

The third scenario generates noisy events compared to the other scenarios and the amount of these events are much more than the action's events, as we observed previously in Figure 5.4. Therefore, we apply the de-noising preprocessing, which is explained in Section 5.4.1, on N-Actions only.

In order to test the NVS domain datasets of all scenarios, a five-fold cross-validation procedure is implemented to find the optimal classifier that can improve the rate of action recognition. Two classifiers, *i.e.*, KNN and QSVM, are used to evaluate the proposed method. KNN classifier is set up with K=1 neighbour.

5.5.2 Evaluation using emulator-based datasets

In this part of the experiment, we evaluate the proposed method on the emulator-based datasets, *i.e.*, E-KTH, E-UCF11, E-UCF50 and E-HMDB50. The frame rates of the corresponding pixel-domain version of these datasets are used to define the size of time interval window, w. Several experiments have been conducted to evaluate the proposed local and global feature extraction, *i.e.*, $\mathcal{F}_T(\mathbb{E})$ and $\mathcal{F}_S(\mathbb{E})$, respectively, as well as the concatenated feature vector, *i.e.*, $\mathbb{F}(\mathbb{E})$. The results of all experiments have been shown in Table 5.2.

The proposed method with concatenated local and global features achieves average accuracy rates of 93.14%, 94.55%, 87.61% for E-KTH, E-UCF11, E-HMDB51, respectively, using QSVM classifier and 69.45% for E-UCF50 using KNN classifier (see Table 5.2). This feature learning improves the accuracy rate of state of the art on the neuromorphic domain for HAR by 0.54%, 19.42% and 25.61% for E-KTH, E-UCF11 and E-HMDB50, respectively. In some

Table 5.2: Accuracy (%) versus the existing work for four datasets: E-KTH, E-UCF11, E-HMDB51 and E-UCF50. These datasets are collected using PIX2NVS emulator.

Method	E-KTH	E-UCF11	E-HMDB51	E-UCF50
Static DVS camera + Hand-crafting	–	75.13 [143]	–	–
Static DVS camera + CNN	92.6 [146]	–	–	–
NVS(emulator) + RGB + CNN	–	–	62.0 [144]	–
NVS(emulator) + Proposed hand-crafting (local only) +KNN	51.17	82.15	73.60	63.66
NVS(emulator) + Proposed Hand-crafting (local only) + QSVM	61.04	76.21	65.41	49.6
NVS(emulator) + Proposed hand-crafting (global only) +KNN	91.47	89.36	73.32	36.62
NVS(emulator) +Proposed Hand-crafting (global only) + QSVM	92.47	92.99	82.82	40.46
NVS(emulator) + Proposed hand-crafting (local-global temporal) +KNN	80.27	93.36	86.38	**69.45**
Proposed Hand-crafting (local-global temporal) + QSVM	**93.14**	**94.55**	**87.61**	65.07

cases, the confusion matrix shows that the classifier cannot discriminate between the actions perfectly. Such cases can be seen in Figure 5.10, when the QSVM classifier recognises between the jogging and running actions because those actions have a similarity. However, the same classifier recognises the boxing action with 98.0% of accuracy.

In general, we conclude that fusing both global and local feature vectors in a single vector outperforms the existing methods on exploring NVS domain in all cases. This concatenated feature vector also achieves the pixel domain-based HAR dataset and improves the accuracy rate of recognition in the RGB version of E-HMDB51 datasets, which is 82.48% [211], by 5.13%. The corresponding confusion matrices of recognizing the actions for these four datasets using QSVM are shown in Figure 5.10, Figure 5.11, Figure 5.12 and Figure 5.13, respectively.

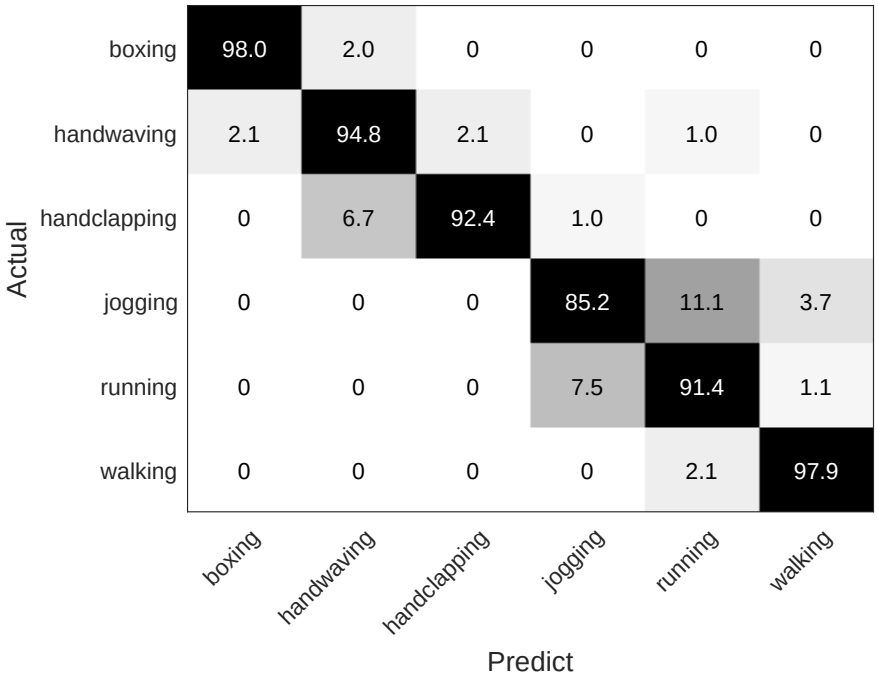


Figure 5.10: Confusion matrix of NVS-based HAR on E-KTH dataset using QSVM (Overall accuracy: 93.14%). The descriptors have been applied on the emulator-based NVS domain.

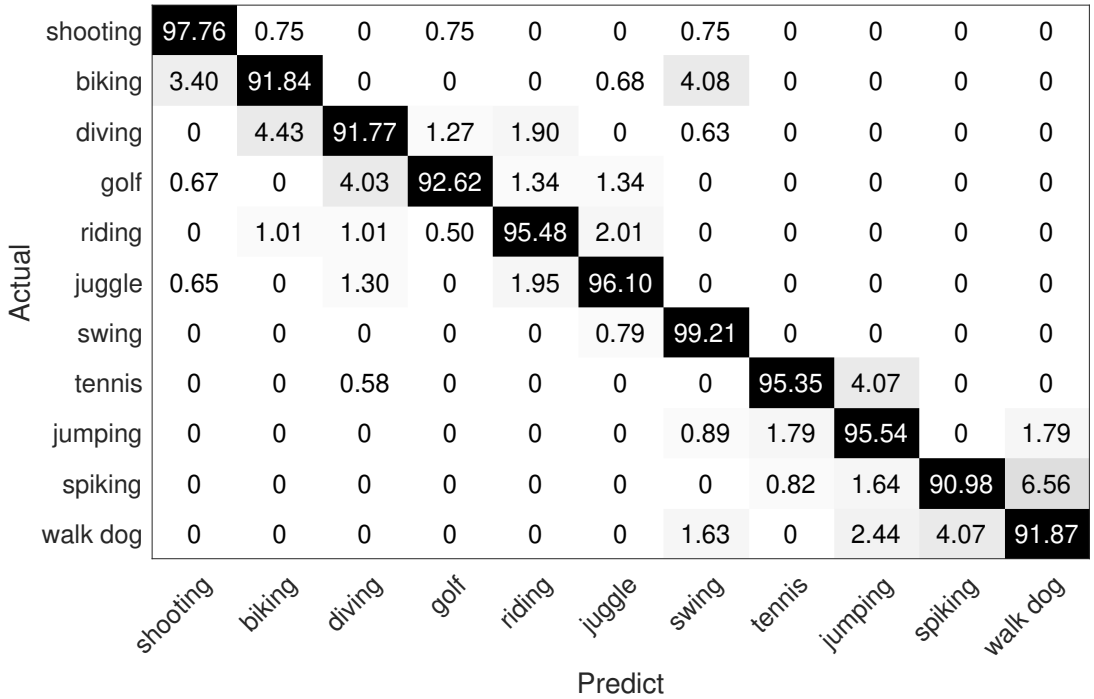


Figure 5.11: Confusion matrix of NVS-based HAR on E-UCF11 dataset (Overall accuracy: 94.43%). The descriptors have been applied on the emulator-based NVS domain.

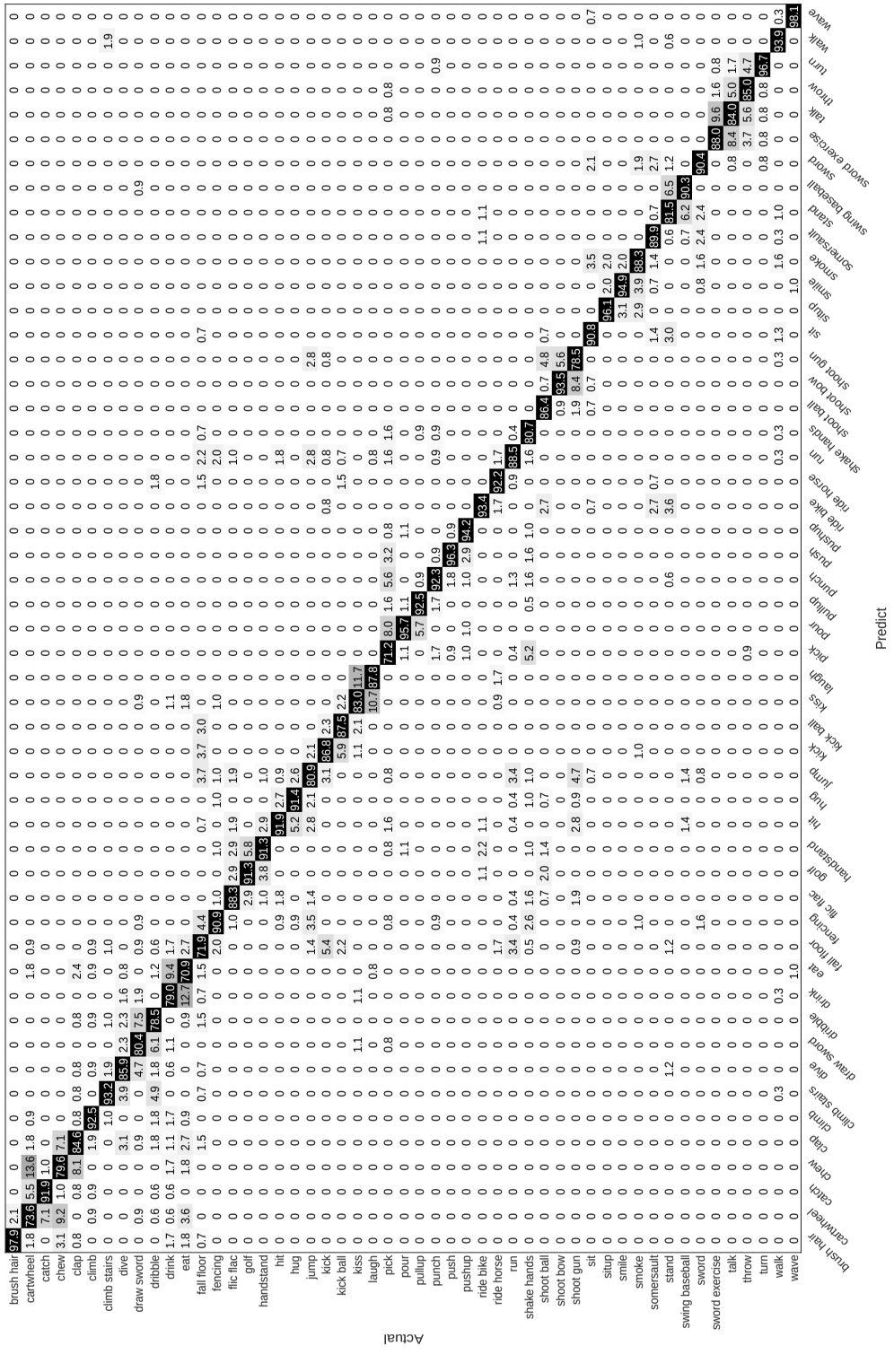


Figure 5.12: Confusion matrix of NVS-based fusion of global and local features on E-HMDB51 dataset using QSVM classifier (Overall accuracy: 87.61%). The descriptors have been applied on the emulator-based NVS domain.

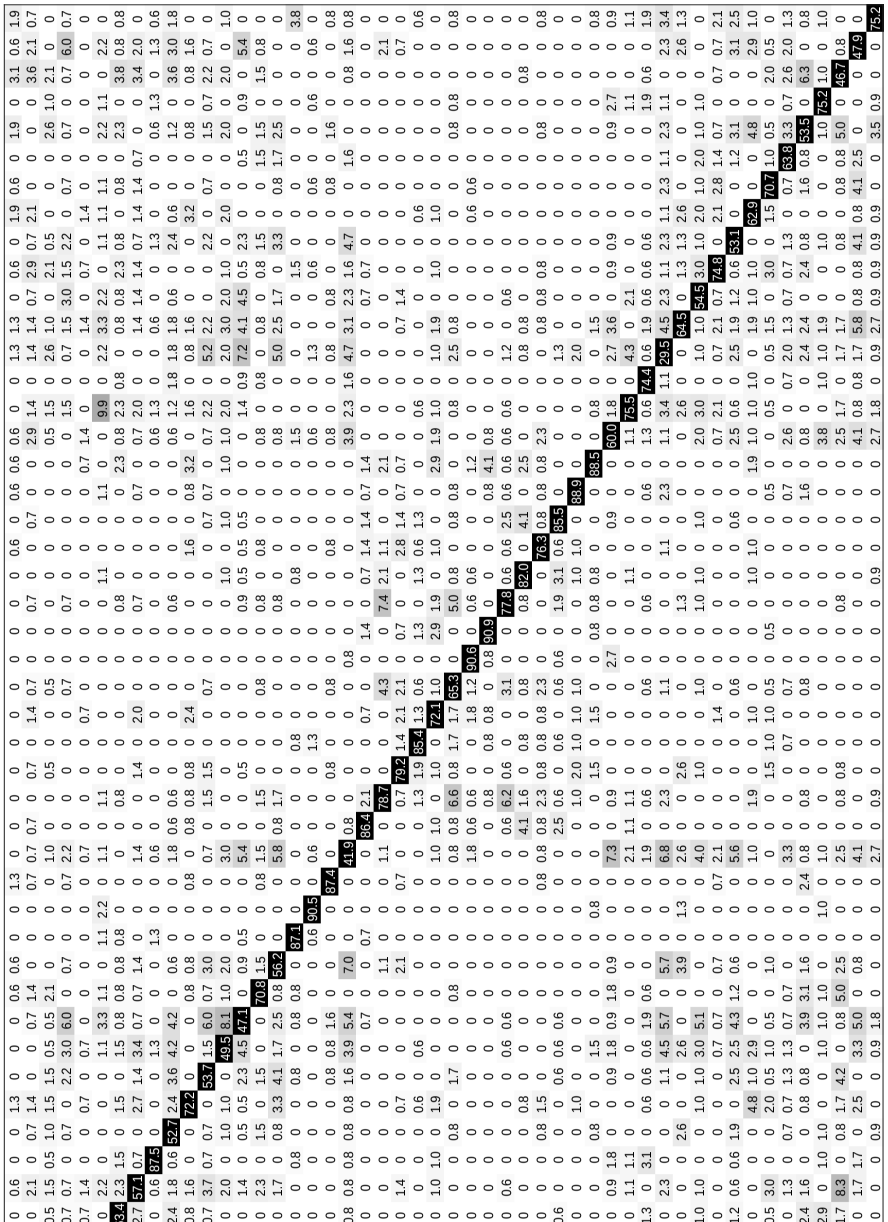


Figure 5.13: Confusion matrix of NVS-based HAR on E-UCF50 dataset (Overall accuracy: 69.81%). The descriptors have been applied on the emulator-based NVS domain.

5.5.3 Evaluation using DVS camera-based datasets

In this part of the experiments, we used two publicly available NVS domain datasets which are acquired using native DVS cameras. The details of these datasets are already explained in Table 5.1. Each one of these datasets is recorded based on a different scenario using two different DVS cameras. In the first scenario, N-Actions dataset in [6] is recorded in an office scene with ten subjects using DAVIS346redColor camera. The second dataset is R-UCF50, which is acquired by playing the original RGB version of UCF50 on the monitor and positioning the DAVIS240C vision sensor camera opposite to the monitor to record the events.

Due to recording these datasets using real DVS cameras, the problem with extracting the local features is in determining the size of the time window. This problem was solved when the emulator is used to generate the event-based version by considering the frame rate of the corresponding pixel-domain version of the dataset to define the size of the window, *i.e.*, w. Thus, we follow the same principle by supposing a $30fps$ to segment the events at each second into 30 time slice windows, $w = 0.03$ seconds.

We started the experiments by applying the de-noising preprocessing step on N-Actions dataset only since the scenario of acquiring this dataset generates a lot of noisy events. These noisy events constitute a large proportion of the total number of events. Thus, removing these events results in a reliable set of features that leads to improve the accuracy rate. The results of the accuracies before and after the de-noising are shown in Table 5.3. As we can see the de-noising improves the accuracy to double because the amount of the noisy events is reduced, and this has been explained in Figure 5.4. We notice that around 70% of the events are noisy and reducing/removing such noisy events improves both the accuracy and the computational complexity of processing the stream of events. To our knowledge, the results in Table 5.3 are the first results on this dataset. The confusion matrix of recognising the actions in N-Actions is shown in Figure 5.14.

In the second scenario based on the DVS camera, R-UCF50 dataset is used to evaluate the proposed method. Despite R-UCF50 being recorded in a controlled environment, we expected that the amount of the noisy events has less effect on the accuracy of recognition. Therefore, the de-noising pre-processing was not applied on R-UCF50. The accuracies of action recognition are illustrated in Table 5.4 and the corresponding confusion matrix is shown in Figure 5.15.

Table 5.3: Accuracy (%) of action recognition using N-Actions: before and after the de-noising.

Method	N-Actions before de-noising	N-Actions after de-noising
DVS camera + Proposed handcrafted (local only)+KNN	38.06	51.21
DVS camera + Proposed handcrafted (local only)+QSVM	33.91	58.48
DVS camera + Proposed handcrafted (global only)+KNN	32.87	44.29
DVS camera + Proposed handcrafted (global only)+QSVM	36.33	43.25
DVS camera + Proposed handcrafted (local-global)+KNN	29.07	53.29
DVS camera + Proposed handcrafted (local-global)+QSVM	37.37	**61.94**

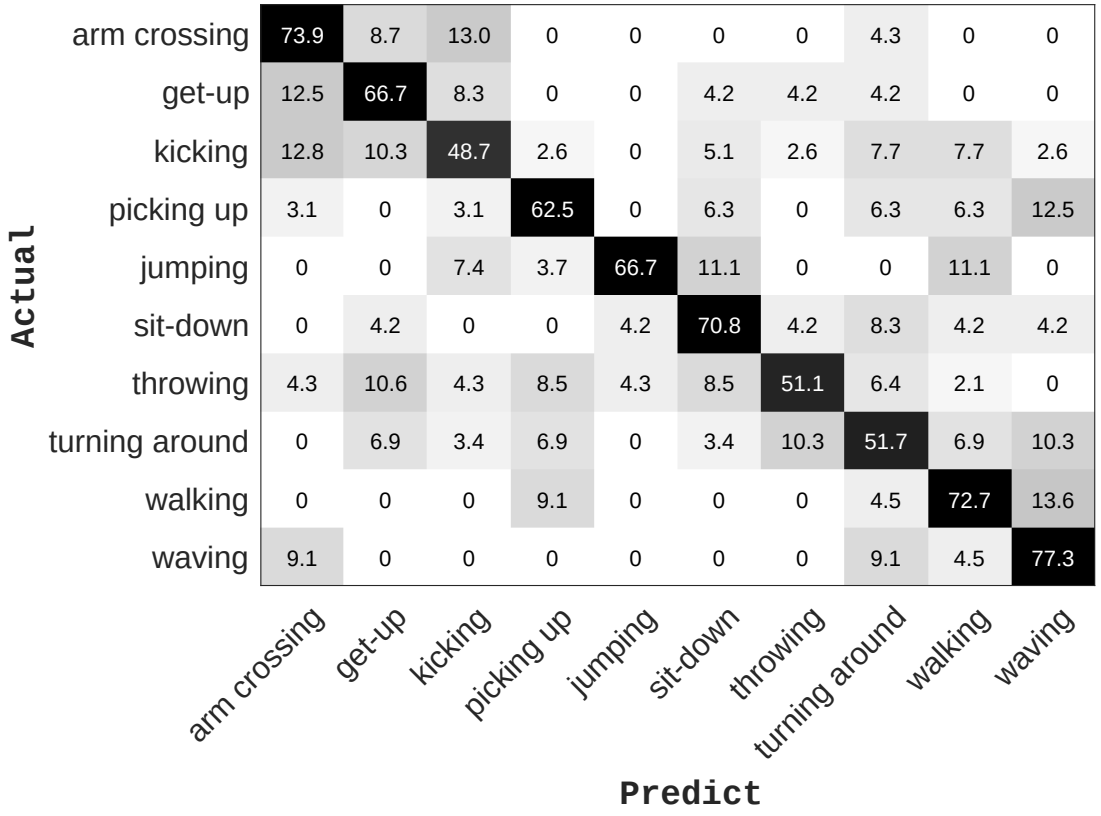


Figure 5.14: Confusion matrix of recognising the actions in [6] using the concatenated feature vectors with QSVM (Overall accuracy: 61.94%). The descriptors have been applied on native NVS domain-based camera after removing the noisy events.

Table 5.4: Accuracy (%) of action recognition of R-UCF50 dataset.

Method	R-UCF50
DVS camera + Proposed handcrafted (local only)+KNN	52.07
DVS camera + Proposed handcrafted (local only)+QSVM	42.49
DVS camera + Proposed handcrafted (global only)+KNN	44.64
DVS camera + Proposed handcrafted (global only)+QSVM	43.6
DVS camera + Proposed handcrafted (local-global)+KNN	**68.96**
DVS camera + Proposed handcrafted (local-global)+QSVM	65.32

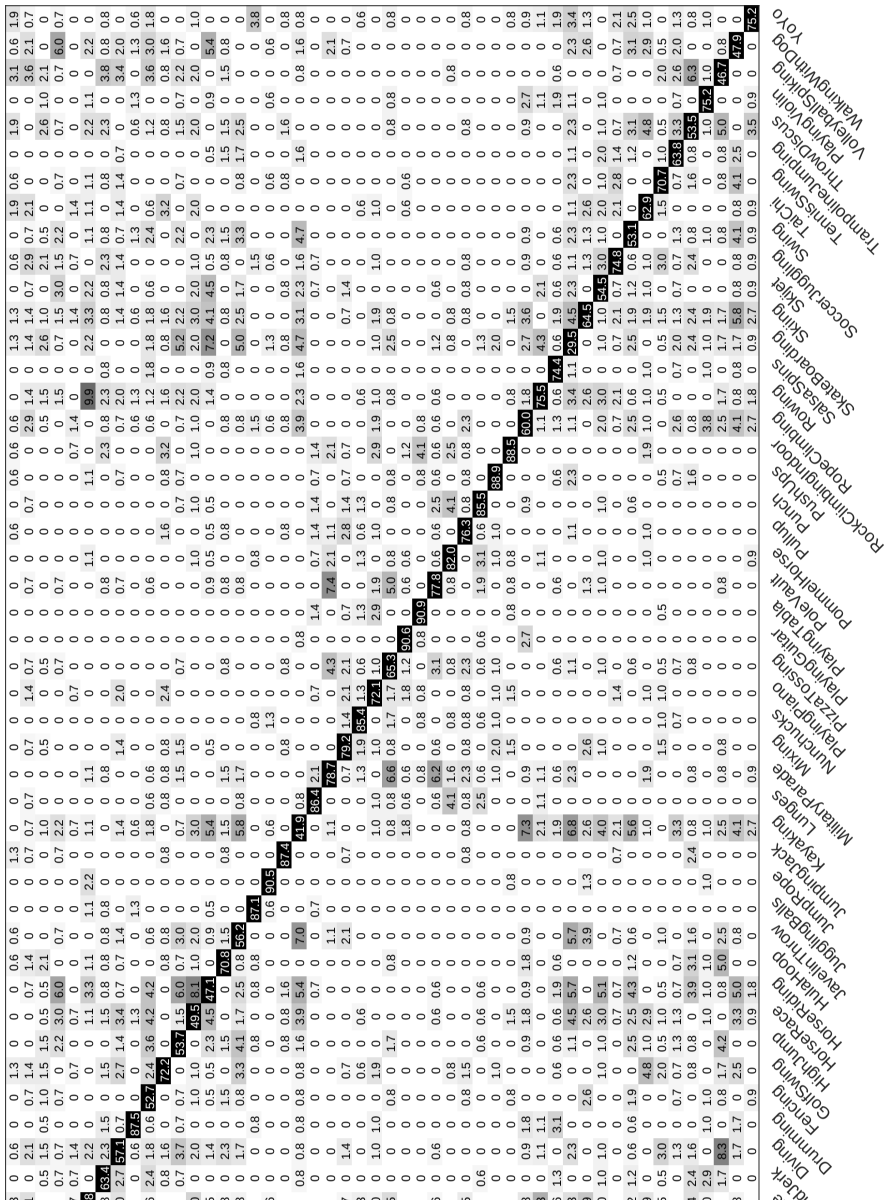


Figure 5.15: Confusion matrix of recognising the actions in R-UCF50 dataset using the concatenated feature vectors with KNN (Overall accuracy: 68.96%).

We observe in Table 5.4 that the accuracy rates of this dataset are less than those generated by the emulator, which are presented in Table 5.2. The reason is that the events generated by the emulator are localised around the objects with a higher density compared to the version that are generated by recording the events using a DVS camera. Another reason is that this scenario seems to generate noisy events more than the emulator-based scenario. These reasons are shown in Figure 5.16, and we notice the localisation of the events based on the emulator in Figure 5.16 (a) compared to those in Figure 5.16(b) and the amount of the noisy events in both figures. However, the de-noising algorithm can be adapted to process the noisy events in R-UCF50 and any dataset that will be recorded using the same scenario to improve the results.

5.5.4 Neuromorphic domain vs. RGB domain: a comparison

We have outlined the accuracy rates of action recognition based on the NVS and RGB domains to compare the performance in the achievement in different domains. The accuracies are collected from Chapter 4 and the current chapter and are shown in Table 5.5 using the RGB, Temporal salience, and NVS versions. The obtained accuracies based on the proposed method are comparable or better than RGB state of the art in most cases. These results make the NVS domain a promising area of research in the applications of privacy protection and HAR.

5.5.5 Computational complexity of the proposed method

All experiments in this chapter were implemented using Matlab R2018a on a PC with Intel processor, CPU@3.6GHz and RAM 16GB. The breakdown of the average times of each step of the algorithm is shown in Table 5.6. This table also shows the computational complexity

Table 5.5: The accuracy rates of action recognition obtained by the RGB, temporal salience, and NVS domains: a comparison.

Dataset	Accuracy (%)		
	RGB	Temporal salience [229]	Proposed NVS
KTH	96.8 [126]	99.06	93.14
UCF11	96.94 [199]	–	94.55
HMDB51	82.48 [211]	99.03	87.61
UCF50	96.4 [204]	–	69.45

Table 5.6: The complexity of the proposed method.

Step	Computational complexity
De-noising	$\mathcal{O}(NC)$
Time slice local feature extraction	$\mathcal{O}(N)$
HRLEP feature extraction	$\mathcal{O}(NH)$
Max RLE	$\mathcal{O}(1)$
Max time	$\mathcal{O}(1)$
Number of ON events	$\mathcal{O}(N)$
Number of OFF events	$\mathcal{O}(N)$
Total	$\mathcal{O}(3N) + \mathcal{O}(NC) + \mathcal{O}(NH) + \mathcal{O}(2)$

of each step. In the step of calculating HRLEP based on RLE algorithm, the average time complexity of run RLE on each event is $\mathcal{O}(N)$, where N is the number of the events in (x, y) over W. Because RLE is run over each event through W, the total time of performing RLE on $\mathbb{E}$ becomes $\mathcal{O}(NH)$. Besides, the pre-processing step of de-noising requires $\mathcal{O}(LC))$, where C is the number of events in each 3D local window on (x, y, t) plane, since this step is applied on each 3D local block. The total computational complexity by including all steps is $\mathcal{O}(3N) + \mathcal{O}(NC) + \mathcal{O}(NH)$. However, the total complexity can be measured by considered $\mathcal{O}(N)$ only.

5.6 Concluding Remarks

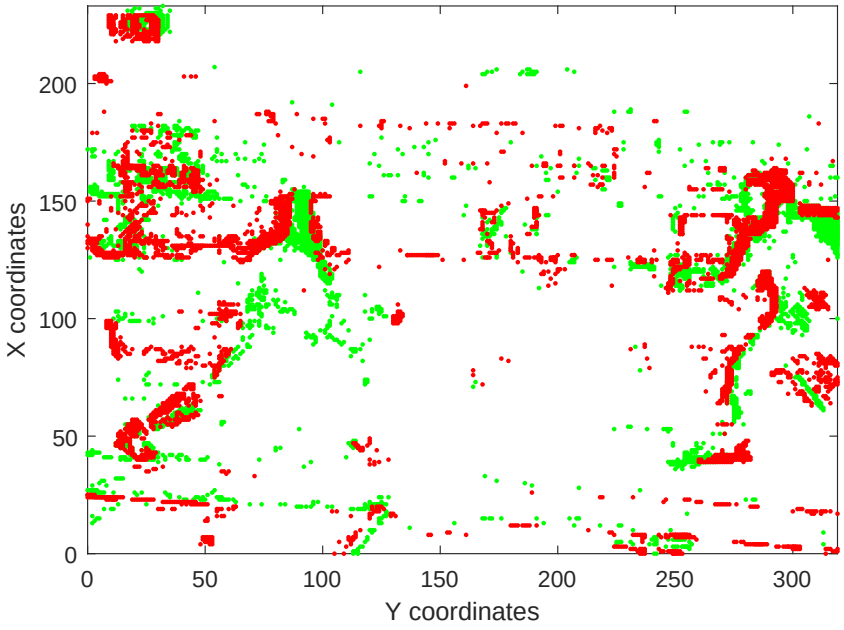
In this chapter, a new de-noising algorithm has been proposed to remove the noisy events that can be recorded, especially in the case of using a native camera for recording the events. Furthermore, two new descriptors exploring the NVS domain for HAR have been proposed. The purpose for proposing these descriptors is to prove that the NVS data satisfies the utility in addition to provide a new tool to anonymise the identity and preserve the privacy in the application of home monitoring, *e.g.*, AAL. The first descriptor extracts locally a set of higher-order descriptive statistics from the events in a time window slice. The second descriptor calculates nine global features by tracking the events along with the whole time interval of the sequence. These descriptors have been tested on several standard datasets. These datasets are categorised into two groups: emulator-based and native DVS camera-based recording. Conducting several experiments using these datasets has proved the reliability of the proposed descriptors to deal with NVS domain and how this domain is worthy of being exploited beyond other computer

vision applications.

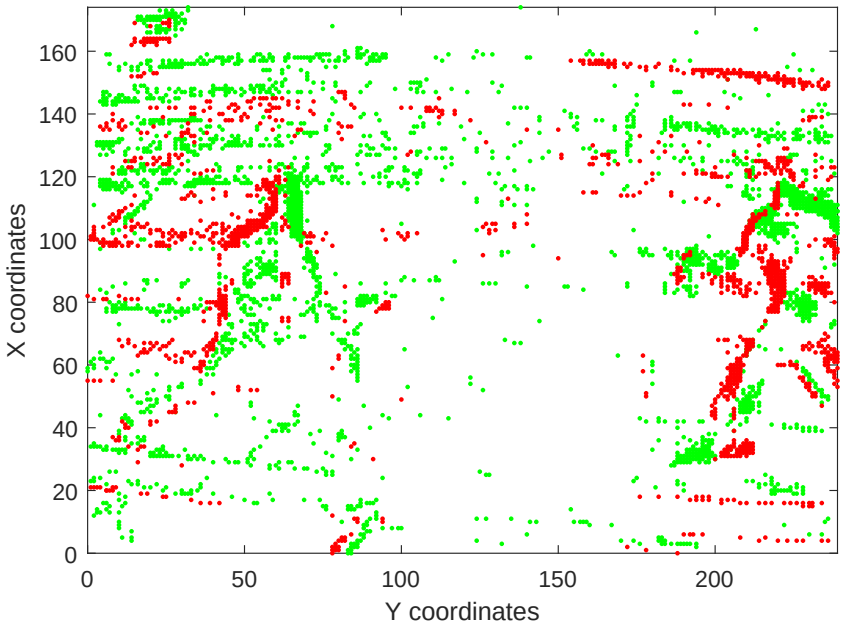
The presented results of the accuracy show the superiority of the proposed descriptor over the existing methods. Demonstrating that the proposed method of exploring the events achieves overall accuracy rates of 93.14%, 94.55%, 87.61%, 69.81%, 61.94% and 68.96% for E-KTH, E-UCF11, E-HMDB51, E-UCF50, N-Actions and R-UCF50 datasets, respectively, outperforming the existing work for all datasets used in the neuromorphic domain. The proposed method improves the accuracy rates by 0.54%, 19.42% and 25.61% for E-KTH, E-UCF11 and E-HMDB50, respectively. The accuracy rates show that the emulator-based NVS domain action recognition achieves the DVS camera-based data. The reason is that the amount of noisy events that are generated from changing the light conditions in the scene when using the DVS camera affects the reliability of extracting the features and then reduces the accuracy of recognition. Besides, applying the de-noising preprocessing on N-Actions dataset duplicates the accuracies, since the number of noisy events is reduced, making the feature vectors more reliable for HAR. Moreover, the proposed method proves its ability to deal with different scenarios of NVS domain-based data. These scenarios include multi-view and real-world video sequences in order to pay attention to increase the interest in using the NVS domain in real-life applications.

However, the proposed de-noising algorithm can be improved to avoid losing action events. Such a case can be noticed in Figure 5.4, whenever, some action's events have been lost from the body of the subject (see Figure 5.4(b)). Another improvement that can be made is by converting the stream of events into frequency domain and extracting frequency-based features, such as, the energy of frequencies. Furthermore, we can apply the graph theory on the stream of events since these events can be considered as points in the spatio-temporal space and find a graph model to link these events. Then, a set of informative features can be extracted from the graph signal.

Lately, deep learning has captured the interest of researchers in various fields. However, the results of applying the deep leaning for HAR based on events are not as good as expected [144, 146]. Therefore, this research area needs more efforts to find out the potential information in the events that can be explored for HAR.



(a) Emulator based extraction



(a) Native DVS camera

Figure 5.16: Two examples for the same frame from a fencing sequence in UCF50 dataset explaining the amount and the distribution of the events in each frame: (a) PIX2NVS emulator has been used to generate the stream of the events and (b) The DVS240C camera has been used to acquire the events. For visualisation, the ON and OFF events are plotted with green and red colours, respectively.

Chapter 6

Conclusions

6.1 Summary of achievements

This thesis has explored the utility of the anonymity domain and addressed the problem of privacy for daily human action recognition in AAL regardless of the trade-off between privacy and utility. Two different anonymity domains have been presented and tested to provide a reliable privacy preservation method. These anonymity domains are categorised into: software-based and device-based, depending on the framework that is used to provide the anonymisation. Each proposed domain takes into account the utility of the obfuscated data for HAR beyond the privacy. We have proposed different methods based on the category of anonymity domain to explore the utility of this domain using a set of publicly available HAR datasets since it is difficult to find real AAL datasets for the reason of ethics.

The first problem was to convert the RGB output into anonymisation form while maintaining data quality. We have presented a new method to preserve the privacy by modelling the action instead of covering the action. The proposed privacy method provides a useful domain that protects privacy, and, at the same time, outputs an action abstraction that can be exploited for action recognition without rendering the original RGB intensities. The proposed method includes a motion detection by computing the difference map between each consecutive frames followed by computing the spectral-based entropy to form temporal saliency map. The salience map contains the obfuscated data by modelling the action silhouette in the frame instead of covering the silhouette. This modelling method creates an anonymisation map based on the

change of the action rather than the structure of the human silhouette. This action modelling has been explained before in Figure 3.4, where different salience maps for different actions are shown. The proposed method has been evaluated using subjective and objective evaluations. These metrics showed that the temporal salience based anonymisation reduces the concern of privacy from the perspective of the individuals. The collected results of the subjective evaluation illustrate that the proposed anonymity domain achieves up to 100% of concealment.

The informative temporal salience map for modelling the action is considered a useful piece of information that is reliable to be used for HAR. The temporal salience based action representation provides the required data of the action model and removes the temporal redundancy of the visual data. The salience data is used to guide the descriptor toward the ROA to extract a new feature, *i.e.*, HOG-S, to train the classifier. The performance of HOG-S improves the accuracy rates of state of the art on most datasets. We observed that the proposed algorithm increases the accuracies of DHA, KTH, UIUC1, UCF sports, and HMDB datasets by 3.04%, 3.14%, 0.83%, 3.67%, and 0.16%, respectively.

The third problem focuses on using a new vision domain, *i.e.*, NVS, for new applications, *e.g.*, anonymisation. The events in the NVS domain represent the intensity changes instead of the intensities without violating privacy as well as outputs useful information that is used for recognising the human actions. The local and global details that are included in the structure of distributing the events spatially and temporally are analysed to extract a new set of features for HAR. The proposed method is applied in two different scenarios, *i.e.*, emulator and native camera. This NVS feature learning performance evaluation showed that the proposed local and global feature vectors improve the accuracies compared to state of the art on E-KTH, E-UCF11, and E-HMDB51 datasets by 0.54%, 19.42% and 25.61%, respectively. In the other scenario, where a native camera is used to acquire the events, the proposed method achieves accuracy rates of 61.94% and 68.96% for N-Actions and R-UCF50 datasets, respectively, and, to our knowledge, these are the first results on these two datasets.

6.2 Future directions

In this section, the presented contributions in this thesis are expanded into some possible future work as follows:

1. In **Chapter 3**, a new method was presented for privacy preservation based on temporal salience detection. This method modelled the human action instead of the human body and omitted the redundant information in the visual data. This approach proves its ability to provide informative abstract that can be used beyond the anonymisation. The proposed method can be developed to provide both anonymity and compression domain at the same time, since it filters the redundancy in the visual data. This compressed and anonymised data, thus, can be transmitted through the cloud without the concern about privacy violation.

2. **Chapter 4**: Since the temporal salience presents a useful abstract and can be used in many applications, the salience abstract can be exploited by the Convolution Neural Network (CNN) for deep feature learning for more semantic tasks, such as activity video understanding and activity video segmentation.

3. In **Chapter 5**, we presented a new method for HAR based on NVS domain events. The proposed method was used for both the anonymisation and feature learning purposes, using the NVS domain. The obtained results indicate the outperforming of the proposed algorithm, however, the achievements are still lower than the results of the corresponding RGB versions. Since the acquired events can be represented in the form of cloud point using three dimensions x, y and $timestamp$, it is easy to formulate these events and their relations as a graph formulation. In this domain, we can apply the presented method in [261] to represent the actions as 3D graph formulation using the relations between the events. This graph modelling can be used to extract local details included in the obtained graph, such as graph spectral features, and used them to recognise the actions since different actions generate different graphs and then different feature vectors.

4. Representing the events as a graph can be applied in many applications. One of the possible applications may include a Graph Convolution Neural Network (GCNN) for classi-

fication or feature representation. Another application is using the graph formulation for graph-based salience silhouette modelling using the sign of the Fiedler vector.

5. Since the neuromorphic domain presents a meaningful motion information for the dynamic objects in the scene, this information seems to be reliable to propose a temporal saliency model based on the NVS domain. The distribution of the saliency can be modelled based on the density of the events in the spatial and temporal coordinates of the stream of events. It may be that exploiting a local graph construction is more productive in this scenario since it links the events in a local region and finds latent details that can be used to form the saliency.

Bibliography

[1] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, J. Kusnitz, M. Debole, S. Esser, T. Delbruck, M. Flickner, and D. Modha, "A low power, fully event-based gesture recognition system," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7243–7252.

[2] H. Kim, Y. Kim, J.-Y. Sim, and C.-S. Kim, "Spatiotemporal saliency detection for video sequences based on random walk with restart," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2552–2564, 2015.

[3] Y. Fang, Z. Wang, W. Lin, and Z. Fang, "Video saliency incorporating spatiotemporal cues and uncertainty weighting," *IEEE Transactions on Image Processing*, vol. 23, no. 9, pp. 3910–3921, 2014.

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

[5] S.-C. Liu and T. Delbruck, "Neuromorphic sensory systems," *Current opinion in neurobiology*, vol. 20, no. 3, pp. 288–295, 2010.

[6] S. Miao, G. Chen, X. Ning, Y. Zi, K. Ren, Z. Bing, and A. Knoll, "Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection," *Frontiers in Neurorobotics*, vol. 13, p. 38, 2019.

[7] F. Palumbo, "Ambient intelligence in assisted living environments," Ph.D. dissertation, 2016.

[8] N. Ryan, T. S. Cinotti, and G. Raffa, "Smart environments and their applications to cultural heritage," *Smart Environments and their Applications to Cultural Heritage*, p. 7, 2005.

[9] D. Lupiana, C. O'Driscoll, and F. Mtenzi, "Taxonomy for ubiquitous computing environments," in *Proceedings of the International Conference on Networked Digital Technologies*, 2009, pp. 469–475.

[10] H. Aloulou, "Framework for ambient assistive living: handling dynamism and uncertainty in real time semantic services provisioning," Ph.D. dissertation, 2013.

[11] C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis, and A. Bauer, "Monitoring activities of daily living in smart homes: Understanding human behavior," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 81–94, 2016.

[12] P. Rashidi and A. Mihailidis, "A survey on ambient-assisted living tools for older adults," *IEEE journal of biomedical and health informatics*, vol. 17, no. 3, pp. 579–590, 2013.

[13] N. C. Krishnan and D. J. Cook, "Activity recognition on streaming sensor data," *Pervasive and mobile computing*, vol. 10, pp. 138–154, 2014.

[14] F. Cardinaux, D. Bhowmik, C. Abhayaratne, and M. Hawley, "Video based technology for ambient assisted living: A review of the literature," *Journal of Ambient Intelligence and Smart Environments*, vol. 3, no. 3, pp. 253–269, 2011.

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

[16] S. Pal and C. Abhayaratne, "Video-based activity level recognition for assisted living using motion features," in *International Conference on Distributed Smart Cameras*, 2015, pp. 62–67.

[17] S. Pal, T. Feng, and C. Abhayaratne, "Real-time recognition of activity levels for ambient assisted living," in *International Conference on Consumer Electronics-Berlin (ICCE-Berlin)*, 2015, pp. 485–488.

[18] S. Pal and C. Abhayaratne, "Phase feature-based activity level estimation for assisted living," in *IET International Conference on Technologies for Active and Assisted Living (TechAAL 2016)*, 2016, pp. 1–6.

[19] T. Delbrück, B. Linares-Barranco, E. Culurciello, and C. Posch, "Activity-driven, event-based vision sensors," in *Proceedings of International Symposium on Circuits and Systems*, 2010, pp. 2426–2429.

[20] T. Delbruck, "Neuromorophic vision sensing and processing," in *European Solid-State Device Research Conference (ESSDERC)*, 2016, pp. 7–14.

[21] V. A. Mateescu, H. Hadizadeh, and I. V. Bajić, "Evaluation of several visual saliency models in terms of gaze prediction accuracy on video," in *Globecom Workshops*, 2012, pp. 1304–1308.

[22] F. W. M. Stentiford, "Visual attention: low-level and high-level viewpoints," in *Optics, Photonics, and Digital Technologies for Multimedia Applications II*, vol. 8436, 2012, p. 84360L.

[23] G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," *arXiv preprint arXiv:1904.08405*, 2019.

[24] Z. Fu, E. Culurciello, P. Lichtsteiner, and T. Delbruck, "Fall detection using an address-event temporal contrast vision sensor," in *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2008, pp. 424–427.

[25] N. B. Bo, F. Deboeverie, M. Eldib, J. Guan, X. Xie, J. Niño, D. Van Haerenborgh, M. Slembrouck, S. Van de Velde, H. Steendam *et al.*, "Human mobility monitoring in very low resolution visual sensor network," *Sensors*, vol. 14, no. 11, pp. 20 800–20 824, 2014.

[26] M. Eldib, F. Deboeverie, W. Philips, and H. Aghajan, "Sleep analysis for elderly care using a low-resolution visual sensor network," in *Human Behavior Understanding*. Springer, 2015, pp. 26–38.

[27] ——, "Behavior analysis for elderly care using a network of low-resolution visual sensors," *Journal of Electronic Imaging*, vol. 25, no. 4, pp. 041 003–041 003, 2016.

[28] R. Li, B. Lu, and K. D. McDonald-Maier, "Cognitive assisted living ambient system: A survey," *Digital Communications and Networks*, vol. 1, no. 4, pp. 229–252, 2015.

[29] H. Foroughi, A. Naseri, A. Saberi, and H. S. Yazdi, "An eigenspace-based approach for human fall detection using integrated time motion image and neural network," in *9th International Conference on Signal Processing (ICSP)*, 2008, pp. 1499–1503.

[30] M. Skubic, G. Alexander, M. Popescu, M. Rantz, and J. Keller, "A smart home application to eldercare: Current status and lessons learned," *Technology and Health Care*, vol. 17, no. 3, pp. 183–201, 2009.

[31] G. Shi, C. S. Chan, W. J. Li, K.-S. Leung, Y. Zou, and Y. Jin, "Mobile human airbag system for fall protection using mems sensors and embedded svm classifier," *Sensors*, vol. 9, no. 5, pp. 495–503, 2009.

[32] S. Bouakaz, M. Vacher, M.-E. B. Chaumon, F. Aman, S. Bekkadja, F. Portet, E. Guillou, S. Rossato, E. Desserée, P. Traineau *et al.*, "CIRDO: Smart companion for helping elderly to live at home for longer," *IRBM*, vol. 35, no. 2, pp. 100–108, 2014.

[33] A. Karpov, L. Akarun, H. Yalçin, A. L. Ronzhin, B. E. Demiröz, A. Çoban, and M. Zeleznỳ, "Audio-visual signal processing in a multimodal assisted living environment." in *INTERSPEECH*, 2014, pp. 1023–1027.

[34] Y. Booranrom, B. Watanapa, and P. Mongkolnam, "Smart bedroom for elderly using kinect," in *International Computer Science and Engineering Conference (ICSEC)*, 2014, pp. 427–432.

[35] M. Kepski and B. Kwolek, "Fall detection using ceiling-mounted 3d depth camera," in *International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 2, 2014, pp. 640–647.

[36] D. Webster and O. Celik, "Systematic review of kinect applications in elderly care and stroke rehabilitation," *Journal of neuroengineering and rehabilitation*, vol. 11, no. 1, p. 1, 2014.

[37] M. Parajuli, D. Tran, W. Ma, and D. Sharma, "Senior health monitoring using kinect," in *Fourth International Conference on Communications and Electronics (ICCE)*, 2012, pp. 309–312.

[38] Š. Obdržálek, G. Kurillo, F. Ofli, R. Bajcsy, E. Seto, H. Jimison, and M. Pavel, "Accuracy and robustness of kinect pose estimation in the context of coaching of elderly population," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2012, pp. 1188–1193.

[39] M. Kepski and B. Kwolek, "Fall detection on embedded platform using kinect and wireless accelerometer," in *International Conference on Computers for Handicapped Persons*. Springer, 2012, pp. 407–414.

[40] C. Zhang, Y. Tian, and E. Capezuti, "Privacy preserving automatic fall detection for elderly using rgbd cameras," in *International Conference on Computers for Handicapped Persons*. Springer, 2012, pp. 625–633.

[41] R. Planinc and M. Kampel, "Introducing the use of depth data for fall detection," *Personal and ubiquitous computing*, vol. 17, no. 6, pp. 1063–1072, 2013.

[42] H.-W. Tzeng, M.-Y. Chen, and J.-Y. Chen, "Design of fall detection system with floor pressure and infrared image," in *International Conference on System Science and Engineering*, 2010, pp. 131–135.

[43] M. V. Sokolova, J. Serrano-Cuerda, J. C. Castillo, and A. Fernández-Caballero, "A fuzzy model for human fall detection in infrared video," *Journal of Intelligent & Fuzzy Systems*, vol. 24, no. 2, pp. 215–228, 2013.

[44] C. Mandel and S. Autexier, "People tracking in ambient assisted living environments using low-cost thermal image cameras," in *International Conference on Smart Homes and Health Telematics*. Springer, 2016, pp. 14–26.

[45] C. Taramasco, T. Rodenas, F. Martinez, P. Fuentes, R. Munoz, R. Olivares, V. H. C. De Albuquerque, and J. Demongeot, "A novel monitoring system for fall detection in older people," *IEEE Access*, vol. 6, pp. 43 563–43 574, 2018.

[46] L. Tao, T. Volonakis, B. Tan, Z. Zhang, and Y. Jing, "3D convolutional neural network for home monitoring using low resolution thermal-sensor array," in *International Conference on Technologies for Active and Assisted Living (TechAAL 2019)*, 2019.

[47] M. T. K. Tsun, B. T. Lau, H. S. Jo, and S. L. Lau, "A human orientation tracking system using template matching and active infrared marker," in *International Conference on Smart Sensors and Application (ICSSA)*, 2015, pp. 116–121.

[48] M. El-hajj, A. Fadlallah, M. Chamoun, and A. Serhrouchni, "A survey of internet of things (IoT) authentication schemes," *Sensors*, vol. 19, no. 5, p. 1141, 2019.

[49] H. Habibzadeh, K. Dinesh, O. R. Shishvan, A. Boggio-Dandry, G. Sharma, and T. Soyata, "A survey of healthcare internet-of-things (HIoT): A clinical perspective," *IEEE Internet of Things Journal*, 2019.

[50] A. Edgcomb and F. Vahid, "Privacy perception and fall detection accuracy for in-home video assistive monitoring with privacy enhancements," *ACM SIGHIT Record*, vol. 2, no. 2, pp. 6–15, 2012.

[51] E. T. Hassan, R. Hasan, P. Shaffer, D. J. Crandall, and A. Kapadia, "Cartooning for enhanced privacy in lifelogging and streaming videos." in *International Conference on Computer Vision and Pattern Recognition Workshops*, Hawaii (USA), 2017, pp. 1333–1342.

[52] C. Ma, A. Shimada, H. Uchiyama, H. Nagahara, and R. Taniguchi, "Fall detection using optical level anonymous image sensing system," *Optics & Laser Technology*, vol. 110, pp. 44–61, 2019.

[53] A. Besmer and H. Lipford, "Tagged photos: concerns, perceptions, and protections," in *Extended Abstracts on Human Factors in Computing Systems*, 2009, pp. 4585–4590.

[54] P. Ilia, I. Polakis, E. Athanasopoulos, F. Maggi, and S. Ioannidis, "Face/off: Preventing privacy leakage from photos in social networks," in *ACM SIGSAC Conference on Computer and Communications Security*, 2015, pp. 781–792.

[55] P. Korshunov, A. Melle, J.-L. Dugelay, and T. Ebrahimi, "Framework for objective evaluation of privacy filters," in *Applications of Digital Image Processing XXXVI*, vol. 8856, 2013, p. 88560T.

[56] A. Li, Q. Li, and W. Gao, "Privacycamera: Cooperative privacy-aware photographing with mobile phones," in *International Conference on Sensing, Communication, and Networking (SECON)*, 2016, pp. 1–9.

[57] N. Vishwamitra, Y. Li, K. Wang, H. Hu, K. Caine, and G.-J. Ahn, "Towards pii-based multiparty access control for photo sharing in online social networks," in *Symposium on Access Control Models and Technologies*, 2017, pp. 155–166.

[58] D. Chen, Y. Chang, R. Yan, and J. Yang, "Protecting personal identification in video," in *Protecting Privacy in Video Surveillance*, 2009, pp. 115–128.

[59] P. Korshunov, C. Araimo, F. De Simone, C. Velardo, J.-L. Dugelay, and T. Ebrahimi, "Subjective study of privacy filters in video surveillance," in *International Workshop on Multimedia Signal Processing (MMSP)*, 2012, pp. 378–382.

[60] B.-J. Han, H. Jeong, and Y.-J. Won, "The privacy protection framework for biometric information in network based CCTV environment," in *International Conference on Open Systems (ICOS)*, Langkawi (Malaysia), 2011, pp. 86–90.

[61] A. Erdélyi, T. Barát, P. Valet, T. Winkler, and B. Rinner, "Adaptive cartooning for privacy protection in camera networks," in *Proceedings of the International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2014, pp. 44–49.

[62] T. Winkler and B. Rinner, "Security and privacy protection in visual sensor networks: A survey," *ACM Computing Surveys (CSUR)*, vol. 47, no. 1, p. 2, 2014.

[63] J. R. Padilla-López, A. A. Chaaraoui, F. Gu, and F. Flórez-Revuelta, "Visual privacy by context: proposal and evaluation of a level-based visualisation scheme," *Sensors*, vol. 15, no. 6, pp. 12 959–12 982, 2015.

[64] A. Pande, P. Mohapatra, and J. Zambreno, "Securing multimedia content using joint compression and encryption," *IEEE MultiMedia*, vol. 20, no. 4, pp. 50–61, 2012.

[65] S. Li, Y. Zhao, B. Qu, and J. Wang, "Image scrambling based on chaotic sequences and veginère cipher," *Multimedia tools and applications*, vol. 66, no. 3, pp. 573–588, 2013.

[66] M.-R. Ra, R. Govindan, and A. Ortega, "P3: Toward privacy-preserving photo sharing," in *USENIX Symposium on Networked Systems Design and Implementation ({NSDI})*, 2013, pp. 515–528.

[67] F. Dufaux, "Video scrambling for privacy protection in video surveillance: recent results and validation framework," in *Mobile Multimedia/Image Processing, Security, and Applications*, vol. 8063, 2011, p. 806302.

[68] Y. Luo, S. C. Sen-ching, R. Lazzeretti, T. Pignata, and M. Barni, "Anonymous subject identification and privacy information management in video surveillance," *International Journal of Information Security*, vol. 17, no. 3, pp. 261–278, 2018.

[69] J. R. Padilla-López, A. A. Chaaraoui, and F. Flórez-Revuelta, "Visual privacy protection methods: A survey," *Expert Systems with Applications*, vol. 42, no. 9, pp. 4177–4195, 2015.

[70] X. Zhang, S.-H. Seo, and C. Wang, "A lightweight encryption method for privacy protection in surveillance videos," *IEEE Access*, vol. 6, pp. 18 074–18 087, 2018.

[71] E. T. Hassan, R. Hasan, P. Shaffer, D. Crandall, and A. Kapadia, "Cartooning for enhanced privacy in lifelogging and streaming videos," in *Internation Conference on Computer Vision and Pattern Recognition Workshops*, 2017, pp. 29–38.

[72] A. A. Chaaraoui, J. R. Padilla-López, F. J. Ferrández-Pastor, M. Nieto-Hidalgo, and F. Flórez-Revuelta, "A vision-based system for intelligent monitoring: human behaviour analysis and privacy by context," *Sensors*, vol. 14, no. 5, pp. 8895–8925, 2014.

[73] W. Q. Yan and F. Liu, "Event analogy based privacy preservation in visual surveillance," in *Image and Video Technology*, 2015, pp. 357–368.

[74] A. Ghanbari and M. Soryani, "Contour-based video inpainting," in *Iranian Conference on Machine Vision and Image Processing*, 2011, pp. 1–5.

[75] A. R. Abraham, A. K. Prabhavathy, and J. D. Shree, "A survey on video inpainting," *International Journal of Computer Applications*, vol. 56, no. 9, 2012.

[76] M. Ebdelli, O. Le Meur, and C. Guillemot, "Video inpainting with short-term windows: application to object removal and error concealment," *IEEE Transactions on Image Processing*, vol. 24, no. 10, pp. 3034–3047, 2015.

[77] S. Çiftçi, A. O. Akyüz, and T. Ebrahimi, "A reliable and reversible image privacy protection based on false colors," *IEEE transactions on Multimedia*, vol. 20, no. 1, pp. 68–81, 2017.

[78] M. Saini, P. K. Atrey, S. Mehrotra, and M. Kankanhalli, "Adaptive transformation for robust privacy protection in video surveillance," *Advances in Multimedia*, vol. 2012, p. 4, 2012.

[79] R. McPherson, R. Shokri, and V. Shmatikov, "Defeating image obfuscation with deep learning," *arXiv preprint arXiv:1609.00408*, 2016.

[80] J. Dai, J. Wu, B. Saghafi, J. Konrad, and P. Ishwar, "Towards privacy-preserving activity recognition using extremely low temporal and spatial resolution cameras," in *International Conference on Computer Vision and Pattern Recognition Workshops*, Boston (USA), 2015, pp. 68–76.

[81] J. Zhao, N. Frumkin, J. Konrad, and P. Ishwar, "Privacy-preserving indoor localization via active scene illumination," in *International Conference on Computer Vision and Pattern Recognition Workshops*, Utah (USA), 2018, pp. 1580–1589.

[82] L. Tao, T. Volonakis, B. Tan, Y. Jing, K. Chetty, and M. Smith, "Home activity monitoring using low resolution infrared sensor array," *arXiv: 1811.05416 v1 [cs. CV]*, pp. 1–8, 2018.

[83] P. Korshunov and T. Ebrahimi, "Towards optimal distortion-based visual privacy filters," in *International Conference on Image Processing (ICIP)*, 2014, pp. 6051–6055.

[84] L. Lorenzen-Huber, M. Boutain, L. J. Camp, K. a. Shankar, and K. H. Connelly, "Privacy, technology, and aging: A proposed framework," *Ageing International*, vol. 36, no. 2, pp. 232–252, 2011.

[85] A. Dimitrievski, E. Zdravevski, P. Lameski, and V. Trajkovik, "Towards application of non-invasive environmental sensors for risks and activity detection," in *IEEE International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2016, pp. 27–33.

[86] L. Atallah, B. Lo, R. Ali, R. King, and G.-Z. Yang, "Real-time activity classification using ambient and wearable sensors," *IEEE Transactions on Information Technology in Biomedicine*, vol. 13, no. 6, pp. 1031–1039, 2009.

[87] N. D. Rodríguez, *Semantic and Fuzzy Modelling for Human Behaviour Recognition in Smart Spaces: A Case Study on Ambient Assisted Living*. IOS Press, 2016, vol. 23.

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

[89] D. J. Patterson, D. Fox, H. Kautz, and M. Philipose, "Fine-grained activity recognition by aggregating abstract object usage," in *IEEE International Symposium on Wearable Computers*, 2005, pp. 44–51.

[90] J. A. Ward, P. Lukowicz, G. Troster, and T. E. Starner, "Activity recognition of assembly tasks using body-worn microphones and accelerometers," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 10, pp. 1553–1567, 2006.

[91] J. Modayil, T. Bai, and H. Kautz, "Improving the recognition of interleaved activities," in *Proceedings of the 10th international conference on Ubiquitous computing*. ACM, 2008, pp. 40–43.

[92] P.-C. Chung and C.-D. Liu, "A daily behavior enabled hidden markov model for human behavior understanding," *Pattern Recognition*, vol. 41, no. 5, pp. 1572–1580, 2008.

[93] C. Nickel, H. Brandt, and C. Busch, "Benchmarking the performance of svms and hmms for accelerometer-based biometric gait recognition," in *IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, 2011, pp. 281–286.

[94] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava, "Using mobile phones to determine transportation modes," *ACM Transactions on Sensor Networks (TOSN)*, vol. 6, no. 2, p. 13, 2010.

[95] D. Trabelsi, S. Mohammed, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "An unsupervised approach for automatic activity recognition based on hidden markov model regression," *IEEE Transactions on Automation Science and Engineering*, vol. 10, no. 3, pp. 829–835, 2013.

[96] F. Martinez-Contreras, C. Orrite-Urunuela, E. Herrero-Jaraba, H. Ragheb, and S. A. Velastin, "Recognizing human actions using silhouette-based hmm," in *IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS'09)*, 2009, pp. 43–48.

[97] L. C. Jatoba, U. Grossmann, C. Kunze, J. Ottenbacher, and W. Stork, "Context-aware mobile health monitoring: Evaluation of different pattern recognition methods for classification of physical activity," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBS)*, 2008, pp. 5250–5253.

[98] M. Shoaib, S. Bosch, O. D. Incel, H. Scholten, and P. J. Havinga, "Fusion of smartphone motion sensors for physical activity recognition," *Sensors*, vol. 14, no. 6, pp. 10 146–10 176, 2014.

[99] A. Wang, G. Chen, J. Yang, S. Zhao, and C.-Y. Chang, "A comparative study on human activity recognition using inertial sensors in a smartphone," *IEEE Sensors Journal*, vol. 16, no. 11, pp. 4566–4578, 2016.

[100] T. van Kasteren and B. Krose, "Bayesian activity recognition in residence for elders," in *3rd IET International Conference on Intelligent Environments (IE 07)*, 2007, pp. 209–212.

[101] P. Rashidi and D. J. Cook, "Keeping the resident in the loop: Adapting the smart home to the user," *IEEE Transactions on systems, man, and cybernetics-part A: systems and humans*, vol. 39, no. 5, pp. 949–959, 2009.

[102] L. Chen, C. D. Nugent, and H. Wang, "A knowledge-driven approach to activity recognition in smart homes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 6, pp. 961–974, 2012.

[103] O. Brdiczka, J. L. Crowley, and P. Reignier, "Learning situation models in a smart home," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 1, pp. 56–63, 2009.

[104] J.-H. Hong, J. Ramos, and A. K. Dey, "Toward personalized activity recognition systems with a semipopulation approach," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 101–112, 2016.

[105] D. L. Vail, M. M. Veloso, and J. D. Lafferty, "Conditional random fields for activity recognition," in *Proceedings of the 6th international joint conference on Autonomous agents and multiagent systems*, 2007, p. 235.

[106] M. Mahdaviani and T. Choudhury, "Fast and scalable training of semi-supervised crfs with application to activity recognition," in *Advances in Neural Information Processing Systems*, 2008, pp. 977–984.

[107] D. H. Hu and Q. Yang, "Cigar: Concurrent and interleaving goal and activity recognition." in *AAAI*, vol. 8, 2008, pp. 1363–1368.

[108] A. M. Khan, Y.-K. Lee, S. Y. Lee, and T.-S. Kim, "A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer," *IEEE transactions on information technology in biomedicine*, vol. 14, no. 5, pp. 1166–1172, 2010.

[109] W. Min, H. Cui, H. Rao, Z. Li, and L. Yao, "Detection of human falls on furniture using scene analysis based on deep learning and activity characteristics," *IEEE Access*, vol. 6, pp. 9324–9335, 2018.

[110] G. Ercolano, D. Riccio, and S. Rossi, "Two deep approaches for adl recognition: A multi-scale lstm and a cnn-lstm with a 3d matrix skeleton representation," in *International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2017, pp. 877–882.

[111] Z. Zhang, X. Ma, H. Wu, and Y. Li, "Fall detection in videos with trajectory-weighted deep-convolutional rank-pooling descriptor," *IEEE Access*, vol. 7, pp. 4135–4144, 2018.

[112] A. Rege, S. Mehra, A. Vann, and Z. Luo, "Vision-based approach to senior healthcare: Depth-based activity recognition with convolutional neural networks," *Semantic Scholar*, 2017.

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

[114] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *International Conference on Computer Vision (ICCV'05)*, 2005, pp. 1395–1402.

[115] A. Veeraraghavan, R. Chellappa, and A. K. Roy-Chowdhury, "The function space of an activity," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2006, pp. 959–968.

[116] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer vision and image understanding*, vol. 104, no. 2-3, pp. 249–257, 2006.

[117] G. Xu and F. Huang, "Viewpoint insensitive action recognition using envelop shape," in *Asian Conference on Computer Vision*, 2007, pp. 477–486.

[118] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, 2007.

[119] D. Tran and A. Sorokin, "Human activity recognition with metric learning," in *European conference on computer vision*, 2008, pp. 548–561.

[120] P. Natarajan and R. Nevatia, "View and scale invariant action recognition using multiview shape-flow models," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

[121] A. Chaaraoui, J. Padilla-Lopez, and F. Flórez-Revuelta, "Fusion of skeletal and silhouette-based features for human action recognition with rgb-d devices," in *Proceedings of the IEEE international conference on computer vision workshops*, 2013, pp. 91–97.

[122] Z. Yin and R. T. Collins, "Shape constrained figure-ground segmentation and tracking," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 731–738.

[123] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using kinect sensor," *IEEE transactions on multimedia*, vol. 15, no. 5, pp. 1110–1120, 2013.

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

[125] Y. Li, W. Li, V. Mahadevan, and N. Vasconcelos, "Vlad3: Encoding dynamics of deep features for action recognition," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 1951–1960.

[126] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream CNN," *IEEE Transactions on Multimedia*, vol. 19, no. 7, pp. 1510–1520, 2017.

[127] S. Kumari and S. K. Mitra, "Human action recognition using DFT," in *National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, 2011, pp. 239–242.

[128] L. Liu, L. Shao, X. Zhen, and X. Li, "Learning discriminative key poses for action recognition," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1860–1870, 2013.

[129] H. Lu, G. Fang, X. Shao, and X. Li, "Segmenting human from photo images based on a coarse-to-fine scheme," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 3, pp. 889–899, 2012.

[130] Y. Ke, R. Sukthankar, and M. Hebert, "Efficient visual event detection using volumetric features," in *International Conference on Computer Vision (ICCV).*, vol. 1, 2005, pp. 166–173.

[131] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM international conference on Multimedia*, 2007, pp. 357–360.

[132] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *European conference on computer vision*, 2008, pp. 650–663.

[133] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–8.

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

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

[136] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *International Conference on Computer Vision (ICCV).*, 2013, pp. 3551–3558.

[137] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Computer Vision and Image Understanding*, vol. 150, pp. 109–125, 2016.

[138] Z. Lan, M. Lin, X. Li, A. G. Hauptmann, and B. Raj, "Beyond gaussian pyramid: Multi-skip feature stacking for action recognition," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 204–212.

[139] S. J. Berlin and M. John, "Human interaction recognition through deep learning network," in *International Carnahan Conference on Security Technology (ICCST)*, 2016, pp. 1–4.

[140] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proceedings of British Machine Vision Conference (BMVC)*, 2008, pp. 275–1.

[141] S. Zhang, Z. Wei, J. Nie, L. Huang, S. Wang, and Z. Li, "A review on human activity recognition using vision-based method," *Journal of healthcare engineering*, vol. 2017, 2017.

[142] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *European conference on computer vision*, 2010, pp. 143–156.

[143] S. A. Baby, B. Vinod, C. Chinni, and K. Mitra, "Dynamic vision sensors for human activity recognition," in *IAPR Asian Conference on Pattern Recognition (ACPR)*, 2017, pp. 316–321.

[144] A. Chadha, Y. Bi, A. Abbas, and Y. Andreopoulos, "Neuromorphic vision sensing for CNN-based action recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 7968–7972.

[145] Y. Bi, A. Chadha, A. Abbas, E. Bourtsoulatze, and Y. Andreopoulos, "Graph-based object classification for neuromorphic vision sensing," in *IEEE International Conference on Computer Vision*, 2019, pp. 491–501.

[146] K. Sullivan and W. Lawson, "Representing motion information from event-based cameras," in *International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2017, pp. 1465–1470.

[147] J. H. Lee, P. K. J. Park, C.-W. Shin, H. Ryu, B. C. Kang, and T. Delbruck, "Touchless hand gesture UI with instantaneous responses," in *International Conference on Image Processing (ICIP)*, 2012, pp. 1957–1960.

[148] A. N. Belbachir, S. Schraml, and A. Nowakowska, "Event-driven stereo vision for fall detection," in *Proceedings of Computer Vision and Pattern Recognition Workshops*, 2011, pp. 78–83.

[149] M. Mahowald, "VLSI analogs of neuronal visual processing: a synthesis of form and function," Ph.D. dissertation, 1992.

[150] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128×128 120 dB 15μs latency asynchronous temporal contrast vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[151] C. Posch, D. Matolin, and R. Wohlgenannt, "A QVGA 143 dB dynamic range frame-free PWM image sensor with lossless pixellevel video compression and time-domain CDS," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 259–275, 2011.

[152] C. Brandli, R. Berner, M. Yang, S.-C. Liu, and T. Delbruck, "A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, 2014.

[153] B. Son, Y. Suh, S. Kim, H. Jung, J.-S. Kim, C. Shin, K. Park, K. Lee, J. Park, J. Woo *et al.*, "A 640× 480 dynamic vision sensor with a 9μm pixel and 300Meps address-event representation," in *International Solid-State Circuits Conference (ISSCC)*, 2017, pp. 66–67.

[154] M. Guo, J. Huang, and S. Chen, "Live demonstration: A 768× 640 pixels 200Meps dynamic vision sensor," in *International Symposium on Circuits and Systems (ISCAS)*, 2017, pp. 1–1.

[155] M. L. Katz, K. Nikolic, and T. Delbruck, "Live demonstration: Behavioural emulation of event-based vision sensors," in *International Symposium on Circuits and Systems*, 2012, pp. 736–740.

[156] T. Delbruck, "Frame-free dynamic digital vision," in *Proceedings of Intl. Symp. on Secure-Life Electronics, Advanced Electronics for Quality Life and Society*, 2008, pp. 21–26.

[157] E. Mueggler, H. Rebecq, G. Gallego, T. Delbruck, and D. Scaramuzza, "The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam," *The International Journal of Robotics Research*, vol. 36, no. 2, pp. 142–149, 2017.

[158] W. Li, S. Saeedi, J. McCormac, R. Clark, D. Tzoumanikas, Q. Ye, Y. Huang, R. Tang, and S. Leutenegger, "Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset," in *British Machine Vision Conference (BMVC)*, 2018, p. 77.

[159] H. Rebecq, D. Gehrig, and D. Scaramuzza, "Esim: an open event camera simulator," in *Conference on Robot Learning*, 2018, pp. 969–982.

[160] G. P. García, P. Camilleri, Q. Liu, and S. Furber, "pyDVS: An extensible, real-time dynamic vision sensor emulator using off-the-shelf hardware," in *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2016, pp. 1–7.

[161] Y. Bi and Y. Andreopoulos, "PIX2NVS: Parameterized conversion of pixel-domain video frames to neuromorphic vision streams," in *International Conference on Image Processing (ICIP)*, 2017, pp. 1990–1994.

[162] J. H. Lee, T. Delbruck, M. Pfeiffer, P. K. J. Park, C.-W. Shin, H. Ryu, and B. C. Kang, "Real-time gesture interface based on event-driven processing from stereo silicon retinas," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 12, pp. 2250–2263, 2014.

[163] T. Serrano-Gotarredona, B. Linares-Barranco, F. Galluppi, L. Plana, and S. Furber, "ConvNets experiments on SpiNNaker," in *International Symposium on Circuits and Systems (ISCAS)*, 2015, pp. 2405–2408.

[164] S. B. Shrestha and G. Orchard, "SLAYER: Spike layer error reassignment in time," in *Advances in Neural Information Processing Systems*, 2018, pp. 1412–1421.

[165] Q. Wang, Y. Zhang, J. Yuan, and Y. Lu, "Space-time event clouds for gesture recognition: From RGB cameras to event cameras," in *Winter Conference on Applications of Computer Vision (WACV)*, 2019, pp. 1826–1835.

[166] G. Chen, J. Chen, M. Lienen, J. Conradt, F. Roehrbein, and A. C. Knoll, "FLGR: Fixed length gists representation learning for RNN-HMM hybrid-based neuromorphic continuous gesture recognition," *Frontiers in neuroscience*, vol. 13, 2019.

[167] A. Chadha, "From pixels to spikes: Efficient multimodal learning in the presence of domain shift," Ph.D. dissertation, UCL (University College London), 2019.

[168] Y. Bi, A. Chadha, A. Abbas, E. Bourtsoulatze, and Y. Andreopoulos, "Graph-based spatial-temporal feature learning for neuromorphic vision sensing," *arXiv preprint arXiv:1910.03579*, 2019.

[169] N. Ikizler-Cinbis and S. Sclaroff, "Web-based classifiers for human action recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 1031–1045, 2012.

[170] D. Wu and L. Shao, "Silhouette analysis-based action recognition via exploiting human poses," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 2, pp. 236–243, 2013.

[171] V. Veeriah, N. Zhuang, and G.-J. Qi, "Differential recurrent neural networks for action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4041–4049.

[172] G. K. Yadav, P. Shukla, and A. Sethfi, "Action recognition using interest points capturing differential motion information," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 1881–1885.

[173] D. Weinland and E. Boyer, "Action recognition using exemplar-based embedding," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1–7.

[174] A.-P. Ta, C. Wolf, G. Lavoue, A. Baskurt, and J.-M. Jolion, "Pairwise features for human action recognition," in *International Conference on Pattern Recognition (ICPR)*, 2010, pp. 3224–3227.

[175] Y. Xie, H. Chang, Z. Li, L. Liang, X. Chen, and D. Zhao, "A unified framework for locating and recognizing human actions," in *International Conference on Computer Vision and Pattern Recognition (CVPR).*, 2011, pp. 25–32.

[176] T. Zhang, L. Xu, J. Yang, P. Shi, and W. Jia, "Sparse coding-based spatiotemporal saliency for action recognition," in *IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 2045–2049.

[177] S. Zeng, G. Lu, and P. Yan, "Enhancing human action recognition via structural average curves analysis," *Signal, Image and Video Processing*, pp. 1–8, 2018.

[178] K. Xu, X. Jiang, and T. Sun, "Two-stream dictionary learning architecture for action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 27, no. 3, pp. 567–576, 2017.

[179] M. Rodriguez, C. Orrite, C. Medrano, and D. Makris, "One-shot learning of human activity with an MAP adapted GMM and simplex-HMM," *IEEE Transactions on Cybernetics*, vol. 47, no. 7, pp. 1769–1780, 2017.

[180] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition." in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, no. 1, 2008, p. 6.

[181] S. O'Hara and B. A. Draper, "Scalable action recognition with a subspace forest," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1210–1217.

[182] L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal laplacian pyramid coding for action recognition," *IEEE Transactions on Cybernetics*, vol. 44, no. 6, pp. 817–827, 2013.

[183] K. Soomro and A. R. Zamir, "Action recognition in realistic sports videos," in *Computer vision in sports*, 2014, pp. 181–208.

[184] T. Wang, Y. Chen, M. Zhang, J. Chen, and H. Snoussi, "Internal transfer learning for improving performance in human action recognition for small datasets," *IEEE Access*, vol. 5, pp. 17 627–17 633, 2017.

[185] A. Ghodrati, A. Diba, M. Pedersoli, T. Tuytelaars, and L. Van Gool, "Deepproposals: Hunting objects and actions by cascading deep convolutional layers," *International Journal of Computer Vision*, vol. 124, no. 2, pp. 115–131, 2017.

[186] L. Wang, Y. Xu, J. Cheng, H. Xia, J. Yin, and J. Wu, "Human action recognition by learning spatio-temporal features with deep neural networks," *IEEE Access*, vol. 6, pp. 17 913–17 922, 2018.

[187] M. H. Siddiqi, M. Alruwaili, A. Ali, S. Alanazi, and F. Zeshan, "Human activity recognition using gaussian mixture hidden conditional random fields," *Computational Intelligence and Neuroscience*, vol. 2019, 2019.

[188] D. Parikh and K. Grauman, "Relative attributes," in *International Conference on Computer Vision (ICCV)*, 2011, pp. 503–510.

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

[190] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu, "Robust relative attributes for human action recognition," *Pattern Analysis and Applications*, vol. 18, no. 1, pp. 157–171, 2015.

[191] Y. Shan, Z. Zhang, P. Yang, and K. Huang, "Adaptive slice representation for human action classification," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 25, no. 10, pp. 1624–1636, 2015.

[192] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[193] I. Everts, J. C. Van Gemert, and T. Gevers, "Evaluation of color stips for human action recognition," in *Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2850–2857.

[194] N. H. Do and K. Yanai, "Automatic construction of action datasets using web videos with density-based cluster analysis and outlier detection," in *Image and Video Technology*, 2015, pp. 160–172.

[195] J. Edwards, M.and Deng and X. Xie, "From pose to activity: Surveying datasets and introducing converse," *Computer Vision and Image Understanding*, vol. 144, pp. 73–105, 2016.

[196] H. Gammulle, S. Denman, S. Sridharan, and C. Fookes, "Two stream lstm: A deep fusion framework for human action recognition," in *Winter Conference on Applications of Computer Vision (WACV)*, 2017, pp. 177–186.

[197] S. Nazir, M. H. Yousaf, J.-C. Nebel, and S. A. Velastin, "A bag of expression framework for improved human action recognition," *Pattern Recognition Letters*, vol. 103, pp. 39–45, 2018.

[198] M. Dai and A. Srivastava, "Video-based action recognition using dimension reduction of deep covariance trajectories," in *Internation Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[199] S. Nazir, M. H. Yousaf, J.-C. Nebel, and S. A. Velastin, "Dynamic spatio-temporal bag of expressions (D-STBoE) model for human action recognition," *Sensors*, vol. 19, no. 12, p. 2790, 2019.

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

[201] Q. Qiu, Z. Jiang, and R. Chellappa, "Sparse dictionary-based representation and recognition of action attributes," in *International Conference on Computer Vision*, 2011, pp. 707–714.

[202] W. Sultani and I. Saleemi, "Human action recognition across datasets by foreground-weighted histogram decomposition," in *International Conference on Computer Vision and Pattern Recognition*, 2014, pp. 764–771.

[203] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik, and V. Albuquerque, "Activity recognition using temporal optical flow convolutional features and multi-layer lstm," *IEEE Transactions on Industrial Electronics*, 2018.

[204] A. Ullah, K. Muhammad, I. U. Haq, and S. W. Baik, "Action recognition using optimized deep autoencoder and cnn for surveillance data streams of non-stationary environments," *Future Generation Computer Systems*, vol. 96, pp. 386–397, 2019.

[205] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: a large video database for human motion recognition," in *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011, pp. 2556–2563.

[206] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4489–4497.

[207] D. Tran, J. Ray, Z. Shou, S.-F. Chang, and M. Paluri, "Convnet architecture search for spatiotemporal feature learning," *arXiv preprint arXiv:1708.05038*, 2017.

[208] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "Actionvlad: Learning spatio-temporal aggregation for action classification," in *International Conference on Computer Vision and Pattern Recognition*, 2017, pp. 971–980.

[209] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6299–6308.

[210] V. Choutas, P. Weinzaepfel, J. Revaud, and C. Schmid, "Potion: Pose motion representation for action recognition," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7024–7033. [Online]. Available: https://hal.inria.fr/hal-01764222

[211] L. Wang, P. Koniusz, and D. Q. Huynh, "Hallucinating bag-of-words and fisher vector IDT terms for CNN-based action recognition," *arXiv preprint arXiv:1906.05910*, 2019.

[212] Y.-C. Lin, M.-C. Hu, W.-H. Cheng, Y.-H. Hsieh, and H.-M. Chen, "Human action recognition and retrieval using sole depth information," in *International Conference on Multimedia*, Nara (Japan), 2012, pp. 1053–1056.

[213] X. Yang, C. Zhang, and Y. Tian, "Recognizing actions using depth motion maps-based histograms of oriented gradients," in *Proceedings of the international conference on Multimedia*, 2012, pp. 1057–1060.

[214] Z. Gao, H. Zhang, G. P. Xu, and Y. B. Xue, "Multi-perspective and multi-modality joint representation and recognition model for 3D action recognition," *Neurocomputing*, vol. 151, pp. 554–564, 2015.

[215] H. Liu, Q. He, and M. Liu, "Human action recognition using adaptive hierarchical depth motion maps and gabor filter," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 1432–1436.

[216] B. Zhang, Y. Yang, C. Chen, L. Yang, J. Han, and L. Shao, "Action recognition using 3D histograms of texture and a multi-class boosting classifier," *IEEE Transactions on Image Processing*, vol. 26, no. 10, pp. 4648–4660, 2017.

[217] M. Liu, H. Liu, and C. Chen, "3D action recognition using multiscale energy-based global ternary image," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 1824–1838, 2018.

[218] R. Li, B. Lu, and K. D. McDonald-Maier, "Cognitive assisted living ambient system: A survey," *Digital Communications and Networks*, vol. 1, no. 4, pp. 229–252, 2015.

[219] T.-H. Tsai and K.-L. Zhang, "Implementation of intelligent home appliances based on IoT," in *Asia Pacific Conference on Circuits and Systems (APCCAS)*, Jeju (Korea), 2016, pp. 146–148.

[220] T. Banerjee, M. Yefimova, J. M. Keller, M. Skubic, D. L. Woods, and M. Rantz, "Exploratory analysis of older adults' sedentary behavior in the primary living area using kinect depth data," *Journal of Ambient Intelligence and Smart Environments*, vol. 9, no. 2, pp. 163–179, 2017.

[221] Z. Luo, J.-T. Hsieh, N. Balachandar, S. Yeung, G. Pusiol, J. Luxenberg, G. Li, L.-J. Li, N. L. Downing, A. Milstein *et al.*, "Computer vision-based descriptive analytics of seniors' daily activities for long-term health monitoring," *Machine Learning for Healthcare (MLHC)*, 2018.

[222] S. Al-Obaidi and C. Abhayaratne, "Privacy protected recognition of activities of daily living in video," in *International Conference on Technologies for Active and Assisted Living (TechAAL 2019)*, 2019, pp. 1–6.

[223] B. Myagmar, J. Li, and S. Kimura, "Heterogeneous daily living activity learning through domain invariant feature subspace," *IEEE Transactions on Big Data*, 2020.

[224] D. J. Butler, J. Huang, F. Roesner, and M. Cakmak, "The privacy-utility tradeoff for remotely teleoperated robots," in *International Conference on Human-Robot Interaction*, Portland (USA), 2015, pp. 27–34.

[225] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 1, pp. 20–33, 2018.

[226] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Advances in neural information processing systems*, 2009, pp. 681–688.

[227] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on context and shape prior." in *British machine vision conference (BMVC)*, vol. 6, no. 7, 2011, p. 9.

[228] C. Posch, R. Benosman, and R. Etienne-Cummings, "Giving machines humanlike eyes," *IEEE Spectrum*, vol. 52, no. 12, pp. 44–49, 2015.

[229] S. Al-Obaidi and C. Abhayaratne, "Temporal salience based human action recognition," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 2017–2021.

[230] J. Chen, "Privacy-preserving smart-room visual analytics," Ph.D. dissertation, Boston University, 2019.

[231] P. Stoica and R. L. Moses, "Spectral analysis of signals," 2005.

[232] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Information-theoretic model comparison unifies saliency metrics," *Proceedings of the National Academy of Sciences*, vol. 112, no. 52, pp. 16 054–16 059, 2015.

[233] W. Lin, M.-T. Sun, R. Poovendran, and Z. Zhang, "Activity recognition using a combination of category components and local models for video surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1128–1139, 2008.

[234] I. Rodomagoulakis, N. Kardaris, V. Pitsikalis, E. Mavroudi, A. Katsamanis, A. Tsiami, and P. Maragos, "Multimodal human action recognition in assistive human-robot interaction," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 2702–2706.

[235] E. Adeli Mosabbeb, K. Raahemifar, and M. Fathy, "Multi-view human activity recognition in distributed camera sensor networks," *Sensors*, vol. 13, no. 7, pp. 8750–8770, 2013.

[236] I. Fatima, M. Fahim, Y.-K. Lee, and S. Lee, "A unified framework for activity recognition-based behavior analysis and action prediction in smart homes," *Sensors*, vol. 13, no. 2, pp. 2682–2699, 2013.

[237] Y. Han, Y. Yang, Y. Yan, Z. Ma, N. Sebe, and X. Zhou, "Semisupervised feature selection via spline regression for video semantic recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 2, pp. 252–264, 2015.

[238] K. Guo, P. Ishwar, and J. Konrad, "Action recognition from video using feature covariance matrices," *IEEE Transactions on Image Processing*, vol. 22, no. 6, pp. 2479–2494, 2013.

[239] R. Minhas, A. A. Mohammed, and Q. M. J. Wu, "Incremental learning in human action recognition based on snippets," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 11, pp. 1529–1541, 2012.

[240] X. Peng, Y. Qiao, Q. Peng, and X. Qi, "Exploring motion boundary based sampling and spatial-temporal context descriptors for action recognition." in *BMVC*, vol. 20, 2013, pp. 93–96.

[241] L. Wang, Y. Qiao, and X. Tang, "Latent hierarchical model of temporal structure for complex activity classification," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 810–822, 2014.

[242] Y. Ye, X. Yang, and Y. Tian, "Exploring pooling strategies based on idiosyncrasies of spatio-temporal interest points," in *Proceedings of the International Conference on Multimedia Retrieval*, 2015, pp. 339–346.

[243] A.-A. Liu, N. Xu, W.-Z. Nie, Y.-T. Su, Y. Wong, and M. Kankanhalli, "Benchmarking a multimodal and multiview and interactive dataset for human action recognition," *IEEE Transactions on cybernetics*, vol. 47, no. 7, pp. 1781–1794, 2017.

[244] K.-P. Chou, M. Prasad, D. Wu, N. Sharma, D.-L. Li, Y.-F. Lin, M. Blumenstein, W.-C. Lin, and C.-T. Lin, "Robust feature-based automated multi-view human action recognition system," *IEEE Access*, vol. 6, pp. 15 283–15 296, 2018.

[245] J. Uijlings, I. C. Duta, E. Sangineto, and N. Sebe, "Video classification with densely extracted HOG/HOF/MBH features: an evaluation of the accuracy/computational efficiency trade-off," *International Journal of Multimedia Information Retrieval*, vol. 4, no. 1, pp. 33–44, 2015.

[246] L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal laplacian pyramid coding for action recognition," *IEEE Transactions on Cybernetics*, vol. 44, no. 6, pp. 817–827, 2014.

[247] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 10, pp. 1915–1926, 2011.

[248] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2005, pp. 886–893.

[249] F. Liu, X. Xu, S. Qiu, C. Qing, and D. Tao, "Simple to complex transfer learning for action recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 949–960, 2016.

[250] F. Murtaza, M. H. Yousaf, and S. A. Velastin, "Multi-view human action recognition using 2D motion templates based on MHIs and their HOG description," *IET Computer Vision*, vol. 10, no. 7, pp. 758–767, 2016.

[251] D. Weinland, M. Özuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010, pp. 635–648.

[252] F. Angelini, Z. Fu, S. A. Velastin, J. A. Chambers, and S. M. Naqvi, "3D-HOG embedding frameworks for single and multi-viewpoints action recognition based on human silhouettes," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[253] W. Li, J. Qiu, and X. Li, "Visual saliency detection based on gradient contrast and color complexity," in *International Conference on Internet Multimedia Computing and Service*, 2015, p. 42.

[254] K. D. Fischl, G. Tognetti, D. R. Mendat, G. Orchard, J. Rattray, C. Sapsanis, L. F. Campbell, L. Elphage, T. E. Niebur, A. Pasciaroni, V. E. Rennoll, H. Romney, S. Walker, P. O. Pouliquen, and A. G. Andreou, "Neuromorphic self-driving robot with retinomorphic

vision and spike-based processing/closed-loop control," in *Conference on Information Sciences and Systems (CISS)*, 2017, pp. 1–6.

[255] D. P. Moeys, F. Corradi, E. Kerr, P. Vance, G. Das, D. Neil, D. Kerr, and T. Delbrück, "Steering a predator robot using a mixed frame/event-driven convolutional neural network," in *International Conference on Event-based Control, Communication, and Signal Processing (EBCCSP)*, 2016, pp. 1–8.

[256] Y. Hu, H. Liu, M. Pfeiffer, and T. Delbruck, "DVS benchmark datasets for object tracking, action recognition, and object recognition," *Frontiers in neuroscience*, vol. 10, p. 405, 2016.

[257] K. A. Boahen, "A burst-mode word-serial address-event link-i: Transmitter design," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 51, no. 7, pp. 1269–1280, 2004.

[258] S.-C. Liu, T. Delbruck, G. Indiveri, A. Whatley, and R. Douglas, *Event-based neuromorphic systems*. John Wiley & Sons, 2014.

[259] T. Stoffregen, G. Gallego, T. Drummond, L. Kleeman, and D. Scaramuzza, "Event-based motion segmentation by motion compensation," in *International Conference on Computer Vision (ICCV)*, 2019, pp. 7244–7253.

[260] R. Serrano-Gotarredona, M. Oster, P. Lichtsteiner, A. Linares-Barranco, R. Paz-Vicente, F. Gómez-Rodríguez, L. Camuñas-Mesa, R. Berner, M. Rivas-Pérez, T. Delbruck, S.-C. Liu, R. Douglas, P. Hafliger, G. Jimenez-Moreno, A. C. Ballcels, T. Serrano-Gotarredona, A. J. Acosta-Jimenez, and B. Linares-Barranco, "CAVIAR: A 45k neuron, 5m synapse, 12g connects/s aer hardware sensory–processing–learning–actuating system for high-speed visual object recognition and tracking," *IEEE Transactions on Neural networks*, vol. 20, no. 9, pp. 1417–1438, 2009.

[261] B. Alwaely and C. Abhayaratne, "Adaptive graph formulation for 3D shape representation," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 1947–1951.