Novel Methods for Posture-Based Human Action Recognition and Activity Anomaly Detection

by

Federico Angelini

A doctoral thesis submitted in partial fulfilment of the requirements for the award of the degree of Doctor of Philosophy (PhD), from Newcastle University.

July 2020

Intelligent Sensing and Communications,
School of Engineering,
Newcastle University,
Newcastle upon Tyne, UK, NE1 7RU.

CERTIFICATE OF ORIGINALITY

This is to certify that I am responsible for the work submitted in this thesis, that the original work is my own except as specified in acknowledgements or in footnotes, and that neither the thesis nor the original work contained therein has been submitted to this or any other institution for a degree.

..... (Signed)

..... (candidate)

To my loving Annamaria,

who is both my handhold, and my strength.

Abstract

Artificial Intelligence (AI) for Human Action Recognition (HAR) and Human Activity Anomaly Detection (HAAD) is an active and exciting research field. Video-based HAR aims to classify human actions and video-based HAAD aims to detect abnormal human activities within data. However, a human is an extremely complex subject and a non-rigid object in the video, which provides great challenges for Computer Vision and Signal Processing. Relevant applications fields are surveillance and public monitoring, assisted living, robotics, human-to-robot interaction, prosthetics, gaming, video captioning, and sports analysis.

The focus of this thesis is on the posture-related HAR and HAAD. The aim is to design computationally-efficient, machine and deep learning-based HAR and HAAD methods which can run in multiple humans monitoring scenarios.

This thesis firstly contributes two novel 3D Histogram of Oriented Gradient (3D-HOG) driven frameworks for silhouette-based HAR. The 3D-HOG state-of-the-art limitations, e.g. unweighted local body areas based processing and unstable performance over different training rounds, are addressed. The proposed methods achieve more accurate results than the baseline, outperforming the state-of-the-art. Experiments are conducted on publicly available datasets, alongside newly recorded data.

This thesis also contributes a new algorithm for human poses-based HAR. In particular, the proposed human poses-based HAR is among the

first, few, simultaneous attempts which have been conducted at the time. The proposed HAR algorithm, named ActionXPose, is based on Convolutional Neural Networks and Long Short-Term Memory. It turns out to be more reliable and computationally advantageous when compared to human silhouette-based approaches. The ActionXPose's flexibility also allows cross-datasets processing and more robustness to occlusions scenarios. Extensive evaluation on publicly available datasets demonstrates the efficacy of ActionXPose over the state-of-the-art. Moreover, newly recorded data, i.e. Intelligent Sensing Lab Dataset (ISLD), is also contributed and exploited to further test ActionXPose in real-world, non-cooperative scenarios.

The last set of contributions in this thesis regards pose-driven, combined HAR and HAAD algorithms. Motivated by ActionXPose achievements, this thesis contributes a new algorithm to simultaneously extract deep-learning-based features from human-poses, RGB Region of Interests (ROIs) and detected objects positions. The proposed method outperforms the state-of-the-art in both HAR and HAAD. The HAR performance is extensively tested on publicly available datasets, including the contributed ISLD dataset. Moreover, to compensate for the lack of data in the field, this thesis also contributes three new datasets for human-posture and objects-positions related HAAD, i.e. BMbD, M-BMdD and JBMOPbD datasets.

Contents

Statement of Originality

Chapter 3

In this chapter, two new frameworks are contributed for silhouette-based human action recognition, by exploiting 3D-HOG features.

1. 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 International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.

Chapter 4

In this chapter, a novel pose-based human action recognition algorithm is contributed, i.e. ActionXPose. Newly recorded video data are also contributed, i.e. ISLD dataset.

2. F. Angelini, Z. Fu, Y. Long, L. Shao, and S. M. Naqvi, *2D Pose-based Real-time Human Action Recognition with Occlusion-handling*, IEEE Transactions on Multimedia, 2019 (Early Access).

Chapter 5

In this chapter, a novel algorithm for combined human action recognition and anomaly detection is contributed, by leveraging information fusions between human poses and RGB data. Newly video datasets are also contributed, i.e. BMbD, M-BMbD and JBMOPbD.

3. F. Angelini and S. M. Naqvi, *Joint RGB-Pose Based Human Action Recognition for Anomaly Detection Applications*, in International Conference on Information Fusion (FUSION), 2019.

4. F. Angelini, Y. Jiawei, and S. M. Naqvi, *Privacy-Preserving Online Human Behaviour Anomaly Detection Based On Body Movements and*

Acknowledgements

Federico Angelini

February, 2020

My gratitude firstly goes to my main supervisor, Dr Syed Mohsen Naqvi, whose constant support and valuable advice has been crucial throughout my PhD study. Particularly, the critical discussion I had the pleasure to undertake with him made the difference to achieve the milestone steps of this research journey.

My gratitude also goes to my secondary supervisors, Dr Satnam Dlay and Prof Jonathon Chambers, for their valuable and supportive advices. I also wish to thank all the other academics I had the pleasure to collaborate with for their helpful contributions, Dr Yang Long, Prof Ling Shao, and Prof Sergio Velastin. A special thank goes to Dr Kianoush Nazarpour and Dr Matthew Dyson, for the fruitful and honest discussions and the opportunities they kindly offered to me.

I wish to thank *Thales*, my industrial supervisor Angus Johnson and Newcastle University to have financially supported my studies and have provided a great, international, research platform.

I would like to acknowledge the numerous people who gave consent to willingly take part of my experiments and data recording. *Thank you all for being so kind!*

I also would like express my gratitude to my colleagues and dear friends, who made this PhD experience humanly unforgettable. In particular (in alphabetical order), Ali Alameer, Ana Carolina Silveira, Hayfaa T. Hussein, Ishita Gulati, Jiawei Yan, Paul Haigh, Safaa N. Awny, Scott Stainton, and Zeyu Fu. *I owe you a lot, guys! Thank you for being you!*

I wish to say a profound and sincere *Thanks* to the closest people I proudly have in this life. Thanks to my friend Paolo, who, among the other

uncountable things he did for me, literally opened the doors of my PhD studies in the UK. Thanks to my family and friends, back in Italy, who are my first and unconditional supporters. *Thank you all, my dears!*

Last, but *definitely* not the least, my deepest gratitude goes to my beloved Annamaria. She *only* knows what doing this PhD really meant to me. She was there, tirelessly, proactively, and sincerely, always ready to live *with* me any good or bad moment this PhD brought us. *I can honestly tell you... we did it, my love, together!*

List of Acronyms

CNN	Convolutional Neural Network
DL	Deep Learning
FC	Fully Connected Layer
FN	False Negative
FP	False Positive
FPR	False Positive Rate
FPS	Frames-Per-Second
HAAD	Human Activity Anomaly Detection
HAR	Human Action Recognition
HC	Hierarchical Clustering
HL	Hierarchical Learning
HOG	Histogram of Oriented Gradients
ISLD	Intelligent Sensing Lab Dataset
LOAO	Leave-One-Actor-Out
LSTM	Long Short-Term Memory
ML	Machine Learning

MLSTM-FCN Multivariate LSTM and Fully Convolutional Network

N Negative

P Positive

PBB Pose-based Branch

PCA Principal Component Analysis

PL Parallel Learning

RBB RGB-based Branch

RGB Red-Green-Blue, Coloured

RNN Recurrent Neural Network

ROC Receiver Operating Characteristic

ROI Region of Interest

SOM Self-Organising Maps

SPF Seconds-Per-Frame

STV Space-Time-Volume

SVM Support Vector Machine

TN True Negative

TP True Positive

TPR True Positive Rate

3D-HOG 3D-Histogram of Oriented Gradient

List of Symbols

$\|.\|_2$, $\|.\|$	Euclidean norm
$(.)^T$	Transpose operator
$\sum$	Summation
$\prod$	Product
min	Minimum value
max	Max value
sgn	Sign
$\forall$	For all
$\in$	Belonging to
$\subset$	Subset of
$. \setminus .$	Difference between sets
$\mathbb{P}(.)$	Probability density
$\mathbf{E}\{.\}$	Expectation
$\mathbb{R}^n$	n-Dimensional Cartesian Space
$\boldsymbol{x}$	Vector
$Var(.)$	Variance
$\|.\|$	Set Cardinality

Chapter 1

INTRODUCTION

1.1 Motivations and Challenges

New powerful technologies and disciplines such as intelligent sensing and machine learning are nowadays pushed forward by both the technological progress and the needs of modern society, e.g. security, surveillance, and assisted living (Figure 1.1). The UK recent history reports tragic incidents

Figure 1.1: (Top-left) CCTV security room continuously monitoring several areas, where automatic surveillance can be crucial [1]. (Top-right) Footage recorded during a terrorist attack (Kerch, 2018), where automatic surveillance could have saved lives [2]. (Bottom-left) Amazon Echo in a living room, where human action recognition capabilities can greatly improve human-device interaction [3]. (Bottom-right) Fall and hazard automatic detection can greatly improve elderly life and allow fast interventions [4].

such as the London and Manchester attacks which highlight the needs of automatic surveillance systems for public security. From a different perspective, the rapid growth of AI-based assisted living systems, such as Alexa and Amazon Echo, are also opening new fascinating research domains. For example, it is possible to imagine future versions of these devices able to interact with humans by interpreting their gestures, movements and even complex actions, identifying user needs, hazards, or simply understanding the action the user is performing to better tailor user-device experience. Applications in healthcare, robotics and IoT would be straightforward and revolutionary. Motivated by this vision, several companies, industries and governments are investing in these advances. The multinational *Thales Group*, which co-funded this thesis project, or the US government-funded *DARPA Agency* are two representative examples of the private and public institutions commitment on following this trend.

This thesis focuses on advancing the state-of-the-art in human action recognition and anomaly detection, by providing new solutions that can potentially be applied to different domains, such as those mentioned above. In the following subsections, some of the most important challenges associated with the focus of this thesis research are provided.

1.1.1 Domain's Constraints and Non-invasive Sensors

Different applications might restrict the usability of certain types of hardware/software solutions. For example, in gaming, human action recognition can easily benefit from depth sensors such as Microsoft Kinect. However, this expensive sensor has strong limitations in outdoor environments and minimal working range [5]. Therefore, for wider area monitoring, RGB cameras might be the best available sensor, since RGB cameras are generally more economical than depth sensors, with relatively wider working range and wider working conditions. In this thesis, other invasive devices such

as accelerometers and motion trackers are not considered, to minimise the number of additional constraints to the human target.

1.1.2 Non-cooperative Scenarios

In literature, two extreme cases are normally considered for action recognition and anomaly detection, i.e. *posed* and *in-the-wild* cases. The first case occurs when human targets perform actions in front of a camera, taking into account pre-defined motion rules, action viewpoints constrains, and explicitly considering that they are monitored. This case normally represents the test-bed for numerous studies. On the other hand, the in-the-wild case represents the opposite of the posed case. Targets do not know they are monitored, they are free to perform any action, regardless any pre-defined instructions. Moreover, the background might be cluttered, multiple targets may be present at the time, targets occlusions might drastically reduce the available information, and recording conditions might be extremely varied. The in-the-wild case represents the most difficult operative condition to be studied.

Publicly available datasets for human action recognition and anomaly detection show a certain degree of *spontaneity*, which varies between posed and in-the-wild (Figure 1.2). However, it is practically difficult to quantitatively assess spontaneity. Therefore, this thesis, in general, focuses on

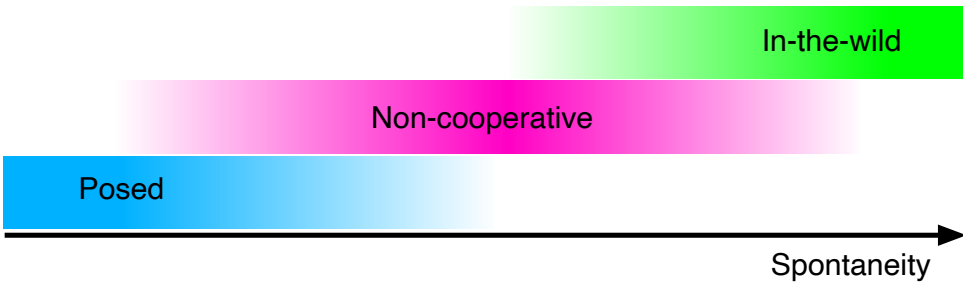


Figure 1.2: Different spontaneity degrees for considered human action recognition and anomaly detection problems. This thesis focuses on non-cooperative scenarios.

non-cooperative scenarios, which can be thought as an half-way between the posed scenarios and the in-the-wild scenarios. However, in few cases, purely posed and in-the-wild datasets will be also considered.

The non-cooperative scenario refers to that condition where the target should not be expected to necessarily collaborate with the model in order to be correctly interpreted. In particular, the targets are aware of being monitored but they might not be cooperative with the data acquisition process. For example, some targets might perform posed actions while other targets might be completely spontaneous. Therefore, spontaneous actions are allowed, leading to action intra-class variations, different action timing, potential occlusions with other targets and objects, viewpoint changes, appearance changes, and contextual information dependence. However, a non-cooperative scenario is still not completely in-the-wild. For example, in-the-wild videos can have cluttered background, moving cameras, noisy frames, extremely dark or bright light conditions, human targets can be severely occluded due to the camera proximity or be undistinguishable from the background, and the number of targets can be extremely high compared to the video resolution, affecting target detection. In contrast, non-cooperative scenarios show mitigated video conditions. In particular, the background is static and not cluttered; the camera is fixed; the video acquisition is not noisy and light condition are ideal; the human targets are fully included within the video frame, despite occlusions might still be visible, and the number of human targets is such that all targets can be detected at all time, except in case of occlusions.

1.1.3 Cross-domain Solutions

A crucial challenge is to define systems that can be applied to different, *unseen* working scenarios with limited or even absent re-calibrations. For example, in object recognition, popular deep learning networks, such as SSD

[10] and YOLO [11], are able to identify objects represented within an RGB image with extremely robust performance. In fact, it is possible to deploy such detectors into domains which are not included within training domains and still obtain good detection performance. In human action recognition and human activity anomaly detection, such cross-domains capabilities are still challenging (Figure 1.3).

1.2 Problem Statements

Formally, Human Action Recognition (HAR) belongs to the machine learning classification problems. Therefore, given some training data describing the performed actions, the goal is to design and train a model to be able to automatically map new input data to a pre-defined set of action labels.

Figure 1.3: Example of cross-domain testing. A model is trained by using multi-domain data [6–9] and tested over a specific domain, depending on the desired application.

In general, classification can be *supervised* (training data is fully labelled), *semi-supervised* (some actions in training data are not labelled or not available), *unsupervised* (training data is not labelled).

In this thesis, regarding HAR, the major focus is on supervised learning, despite other unsupervised techniques have also been used as complementary data processing steps. Regarding the focused human actions, the major focus throughout this thesis is on *posture-related* actions, i.e. those human actions which can be mostly identified by the overall body and limbs movements in the spatio-temporal domain, such as *walking*, *running*, *bending*, *hand-waving*, and *sitting*. More complex actions, such as *drinking*, *riding-bike* or *making-up*, where the targets necessarily interact with objects to perform the action, will play a minor role and will be discussed only in particular experiments.

Similar to HAR, Human Activity Anomaly Detection (HAAD) belongs to the binary classification problems. In other words, the HAAD goal is to exploit the trained model to map new input data to a binary label, i.e. *normal/abnormal*. In this thesis, HAAD is always performed in a semi-supervised fashion, i.e. training data only contain normal instances. This case best simulates the common anomaly detection scenario, where abnormal data is usually missing due to the meagre rate of abnormal events. Therefore, training data define what the model expects to be *normal* and any variations from this expectation will be considered as an abnormality. Similar to HAR, most of the HAAD models will be trained on posture-related normal activities. Additionally, in some experiments, key objects positions in the space-time domain will also be considered as complementary HAAD level of analysis.

1.3 Aims and Objectives

Given the above-mentioned problems and the challenges discussed in Section 1.1, this thesis firstly focuses on proposing human silhouette-based and human pose-driven (Figure 1.4) approaches to perform unimodal human action recognition.

The aim is to obtain robust and computationally-efficient models for HAR and HAAD, which can potentially be deployed in non-cooperative, multiple humans monitoring and surveillance scenarios.

On the basis of the literature review performed throughout this project, the major attention was firstly driven towards human-silhouette based HAR. Therefore, to compensate human-silhouette limitations, the main focus turned on human-poses for HAR and HAAD. Finally, to further compensate for human-poses limitations, advanced models for multimodal RGB-pose features extraction were studied for HAR and HAAD. In table 1.1, the

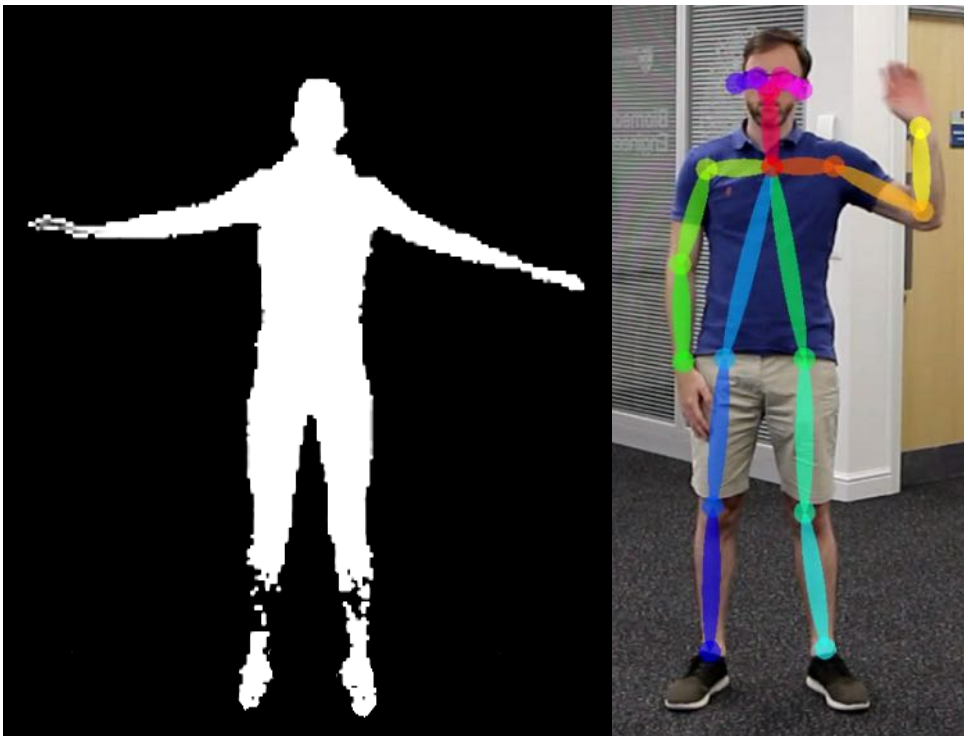


Figure 1.4: (Left) Human silhouette example. (Right) Human pose example.

main features for the methods developed in this thesis are summarised.

Table 1.1: Summary for the developed methods in this thesis.

	Classification	**Classes**	**Action Semantic**	**Scenario**
HAR	supervised	multi-class	posture-related	Posed; Non-cooperative;
HAAD	semi-supervised	binary	posture-related	Non-cooperative;

In Figure 1.5, a visual overview of the whole thesis is provided, which shows the logical connections between chapters and thesis objectives. In particular, the following objectives have been considered:

Objective 1

In Chapter 3, a mechanism to strengthen state-of-the-art silhouettes and 3D-HOG based HAR is investigated, by exploiting local body parts based attention mechanisms.

Objective 2

In Chapter 3, performance stability and robustness over different training rounds is ensured by leveraging cross-actions local gestures data clusters.

Objective 3

In Chapter 3, testing on non-cooperative scenarios is performed, to assess limitations of the human silhouettes-based HAR in terms of reliability for real-world problems. New data recording is required.

Objective 4

In Chapter 4, human silhouettes embedding based mechanism, defined in Chapter 3, are transformed into human pose based embedding, to overcome the human silhouettes limitations.

Objective 5

In Chapter 4, human pose generality is exploited to perform cross-datasets implementations. New data recording is required to test cross-datasets performance.

Objective 6

Chapter 4, based on the human pose detector limitations, i.e. sudden false negative and missing data due to occlusions, new solutions are proposed.

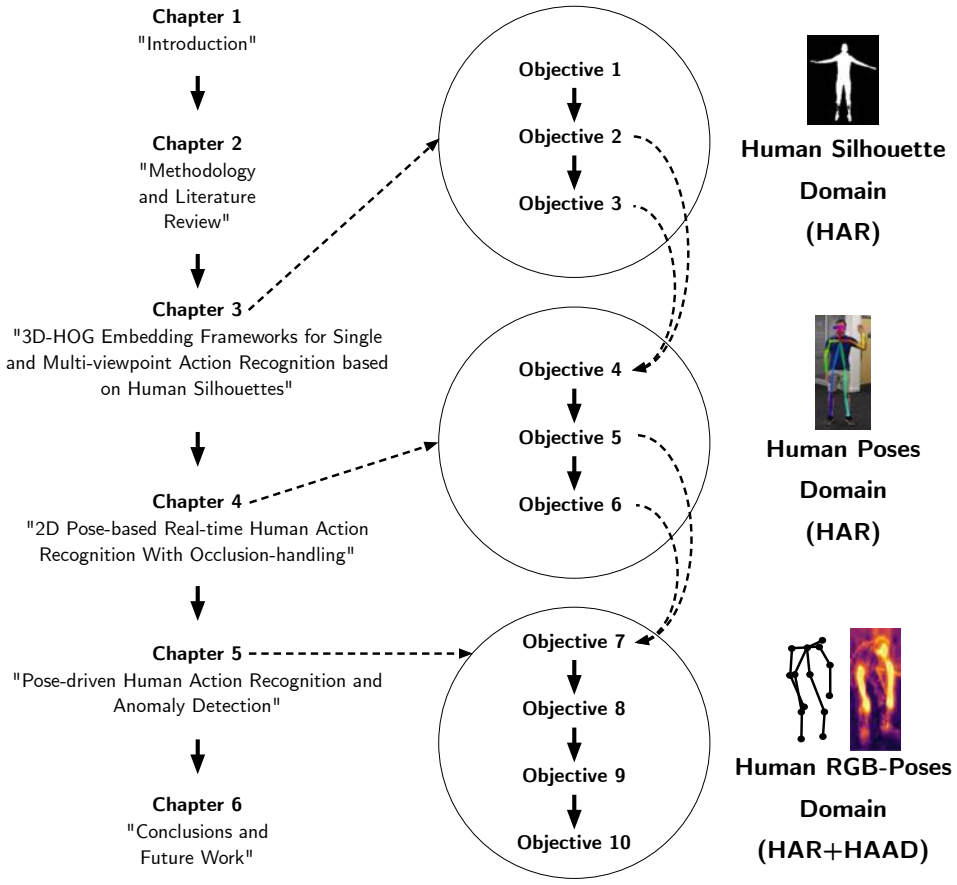


Figure 1.5: Thesis structure. The objectives defined in Section 1.3 are linked following a logical order. The diagram also reports the video-based application domain for each contribution chapter.

Objective 7

In Chapter 5, a generalisation of the pose-based HAR, defined in Chapter 4, to Joint RGB-pose based HAR is introduced, to leverage the additional knowledge carried by RGB data and compensate human pose limitations.

Objective 8

In Chapter 5, the HAAD problem is addressed by considering pose-based and RGB-based deep learning features to train a semi-supervised anomaly detection system. New data is required for non-cooperative HAAD evaluation.

Objective 9

In Chapter 5, object position related anomalies are considered for bilevel body-motion/object-position HAAD, by defining proper object-position features and replicating the Objective 7 approach. New data is required for bilevel and non-cooperative HAAD evaluation.

Objective 10

In Chapter 5, multi-target generalisation is proposed for the novel Pose-driven, bilevel, combined HAR and HAAD. New data recording is required to collect novel, multi-target, non-cooperative HAAD datasets. New data is required for multi-target and non-cooperative HAAD evaluation.

1.3.1 Definitions and Notations

HAR

Let $\mathcal{L} = \{l_i\}_{i=1}^{L}$ be the set of action labels, where L is the total number of considered action labels. Let $\mathcal{W} = \{w_i\}_{i=1}^{W}$ be the set of viewpoints, where W is the total number of considered viewpoints. Therefore, let $\mathbb{D} = \{(s, l, w)_i\}_{i=1}^{D}$ be the multi-viewpoints action dataset containing D samples,

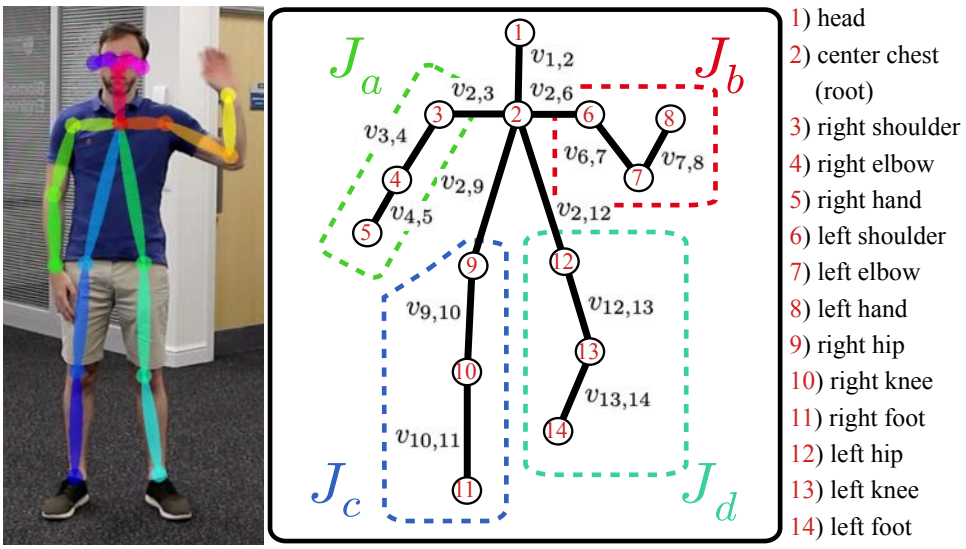


Figure 1.6: The human pose is a set of 2D landmarks J, which represents a non-redundant description of the target posture. (Right) Examples of a human pose. (Left) Proposed pose nomenclature, for pose landmarks $J = \{1, \ldots, 14\}$, link vectors v_{j_1,j_2}, for $j_1, j_2 \in J$, and considered landmark subsets, i.e. J_a, J_b, J_c and J_d.

where s is the RGB video sample, $l \in \mathcal{L}$, and $w \in \mathcal{W}$.

Let $\mathbb{T} \subset \mathbb{D}$ be the chosen training subset and $\mathbb{T}^* = \mathbb{D} \backslash \mathbb{T}$ be the testing subset. In this thesis, HAR aims to extract features from each sequence $s_i \in \mathbb{D}$ to train a model by using $\mathbb{T}$ to recognise actions sequences in $\mathbb{T}^*$.

A human pose detector provides a root-centred graph in the form of 2D coordinates, for example as shown in Figure 1.6, such that

$$p_i(t) = \{(x_j(t), y_j(t))_i\}_{j \in J} \quad t \in \{1, \ldots, T_i\} \tag{1.3.1}$$

where T_i represents the time length of the i-th video sample and J is the landmarks set defined by the pose detector mapping. In this thesis, $J = \{1, \ldots, 14\}$, since the chosen detector is OpenPose [12], whose implemented version detects up to 14 different landmarks. Let v_{j_1,j_2} be the link vector between body landmarks $j_1, j_2 \in J$, as defined in Figure 1.6.

In Chapter 4, the main goal is to exploit $p_i(t)$, as well as generate additional and more robust time sequences, to train a recurrent neural network

using $\mathbb{T}$ and to predict action labels in $\mathbb{T}^*$.

HAAD

In the case of HAAD, let $\mathbb{D} = \mathbb{N} \cup \mathbb{T}$ a video dataset, where $\mathbb{N}$ contains *normal* data while $\mathbb{T}$ represents *testing* data, which contains both normal and abnormal data. In particular, let $\mathbb{N} = \{\mathbf{s}_i\}_{i=1}^{N}$ be the normal video data subset containing N clips of arbitrary time length, where $\mathbf{s}_i$ represents the i-th RGB video clip. Let

$$\mathbb{T} = \left\{ \left(\bar{\mathbf{s}},\ \{g_{\bar{\mathbf{s}}}(t|\phi)\}_{\phi=1}^{d} \right)_i \right\}_{i=1}^{M} \tag{1.3.2}$$

be the video testing dataset containing m clips, where $\bar{\mathbf{s}}$ represents the RGB video clip, $g_{\bar{\mathbf{s}}}(t|\phi)$ is the anomaly based ground truth given for the target identity index $\phi \in \{1, \ldots, \Phi\}$, where Φ is the number of human targets. In particular, $g_{\bar{\mathbf{s}}}(t|\phi)$ is a function that associates each frame t and target identity ϕ to a binary response, i.e. normal/abnormal label, and it can be defined as follows:

$$g_{\bar{\mathbf{s}}}(t|\phi) = \begin{cases} 0 & \text{if } \phi \text{ performs normal action in } t \\ 1 & \text{if } \phi \text{ performs abnormal action in } t \end{cases} \tag{1.3.3}$$

Since $\mathbb{D}$ can be a single or a multi-target dataset, the target identity is provided as discussed in Section 5.2.1 by a tracking mechanism operating over the target detections.

In this thesis, the goal of HAAD is to propose a strategy to compute $G_{\bar{\mathbf{s}}}(t|\phi) \approx g_{\bar{\mathbf{s}}}(t|\phi)$ for all ϕ, and simultaneously provide an explanation of the performed action, defined as $L_{\bar{\mathbf{s}}}(t|\phi)$, in the form of an action label. In principle, action anomalies within $\mathbb{T}$ might be due to the unexpected target's body movements or with an unexpected positioning/usage of contextual objects. Therefore, the aim of this work is to propose an anomaly detection

algorithm that can potentially detect both anomalies.

1.4 Contributions

To the best of my knowledge, and on the basis of the above-mentioned aims and objectives, the main contributions of this thesis to the HAR and HAAD fields are the following:

- Inspired by 3D-HOG previous studies, in Chapter 3, novel, outperforming 3D-HOG frameworks for silhouette-based HAR are contributed;
- Inspired by 3D skeleton literature, in Chapter 4, for the first time, the 2D pose-based HAR is proposed and investigated as a new research area;
- Inspired by information fusion, in Chapter 5, novel, joint RGB-pose based networks are contributed, to simultaneously solve HAR and HAAD non-cooperative problems.
- This thesis also contributes new video datasets, i.e. ISLD-2018, ISLD, and ISLD-2019 for posed and non-cooperative, posture-based HAR. Moreover, this thesis also contributes a novel video dataset ISLD-A (including BMbD, M-BMbD and JBMOPbD datasets), for non-cooperative, multi-target, and posture-based HAAD.

In Chapter 2, further insights about this thesis contributions are provided, including extensive discussions about related works and impact that the proposed techniques have on the HAR and HAAD literature.

1.5 Thesis Outline

In Chapter 2, the main signal processing and computer vision based methods exploited in this thesis are firstly introduced. They represent solid and

well-established methods for classification, clustering, anomaly detection, based on conventional machine learning techniques and deep learning neural networks. Therefore, a literature review for HAR and HAAD is presented, to relate this thesis work to existing works. In Chapter 3, the contributed method for silhouette based HAR is presented. Based on the background subtraction limitations which affected the silhouette based HAR, in Chapter 4, contributions related to human pose-based HAR are provided. Therefore, based on the pose-based HAR limitations, further and conclusive contributions on joint RGB-pose HAR and HAAD are detailed in Chapter 5. In Chapter 6, conclusions, critical discussion and future work is provided.

Chapter 2

METHODOLOGY AND LITERATURE REVIEW

2.1 Methodology Overview

In this thesis, Machine Learning (ML) and Deep Learning (DL) [13] has been exploited as main tools for HAR and HAAD. Modern Artificial Intelligence (AI) rely on sophisticated ML and DL algorithms which can explore training data, finding useful relationships between data features. Such relationships yield to non-linear *rules* which, to some extent, might generalise to new input data. From a different perspective, ML and DL algorithms learn from training data on how to transform new input data into new representations, which are relevant to the given purpose.

In Table 2.1, the relevant ML and DL methods exploited in this thesis are summarised. The purposes of these methods are mainly for dimensionality reduction, classification, anomaly detection and clustering. Dimensionality reduction refers to the ability to reduce data dimensionality by, for example, reducing information redundancy within data or selecting only most relevant features. The classification refers to the ability to classify input data by assigning a single label chosen among a pre-defined set of labels. Similarly, anomaly detection can be considered as a binary classification, where the two labels are *normal* or *abnormal*. Lastly, clustering methods aims to explore

input data in order to highlight internal grouping structures based on a pre-defined motion of distance.

All these methods may belong to the supervised, unsupervised or semi-supervised learning. The supervised learning aims to exploit explicitly labelled instances as training data. It is the case of the supervised classification, where the training aims to learn how extract general rules by exploiting the known one-to-one assignment between training data and their labels. Conversely, unsupervised learning does not exploit any additional knowledge other than input data, e.g. no labels are considered, to train the model. It is the case of clustering, where the group subdivision, i.e. cluster's labels, are the output of the process, instead of part of the input. Lastly, semi-supervised learning may refer to different meanings. In this thesis, semi-supervised learning is only referred to anomaly detection, following the definition given by Goldstein [14]. Semi-supervised anomaly detection is opposed to Unsupervised and Supervised anomaly detection, which are both beyond the scope of this thesis. According to the Goldstein definition, in a semi-supervised anomaly detection model, training data only provides *normal* instances, while *abnormal* instances are unknown. Thus, the trained model will only build up expectation related to normal events, supported by training data. As a consequence, an *unseen* instance which sufficiently differs from training data will be considered as abnormal. Similarly, less frequent actions within training data will also likely to be considered as abnormal instances in the testing phase.

The literature about these ML and DL algorithms is superbly wide and extensive. Therefore, to fit the specific purpose of this thesis, only a brief introduction about relevant methods is provided in the remaining of this section. However, in each contribution chapter, detailed information about the proposed ML and DL algorithm and how it has been designed is provided, including relevant literature references which further support the proposed

solutions.

Table 2.1: Main methods are summarised for a methodology overview. The purpose can be Dimensionality Reduction (DR), Classification (Class), Anomaly Detection (AD) and Clustering (Clus). The training mode can be Supervised (S), Semi-supervised (S-s) and Unsupervised (U).

Method	Purpose	Training
Principal Component Analysis (PCA)	DR	U
Logistic Regression	Class	S
Support Vector Machine (SVM)	Class, AD	S, S-s
CNN/RNN	Class	S
Autoencoders	DR, AD	U, S-s
Hierarchical Clustering	Clus	U
Self-organizing Map (SOM)	DR, Clus	U

2.1.1 Principal Component Analysis (PCA)

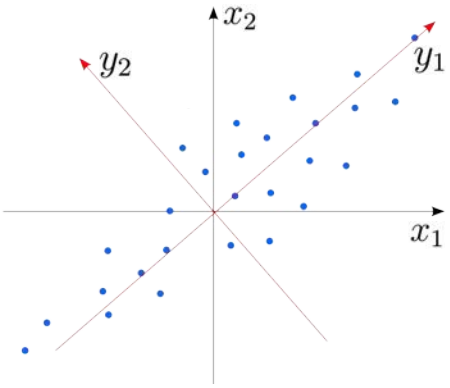


Figure 2.1: PCA transformation example. The zero-mean data $\boldsymbol{x} = [x_1, x_2]$ is transformed into $\boldsymbol{y} = [y_1, y_2]$, which are, respectively, the first and second principal component.

Given a set of zero-mean data in the form of vectors $\boldsymbol{x} \in \mathbb{R}^n$, Principal Component Analysis (PCA) [15] aims to reduce information redundancy within data by mean of a linear transformation learned from training data. Since some of the entries of $\boldsymbol{x}$ might be statistically correlated, PCA aims to linearly transform $\boldsymbol{x}$ into $\boldsymbol{y} \in \mathbb{R}^m$ such that $m < n$, variable in $\boldsymbol{y}$ are uncorrelated to each other and their internal variance is maximised. The goal is achieved by considering $\boldsymbol{x}$ as a multivariate random vector and the

entries of $\boldsymbol{y}$, i.e. y_p, are expected to be such that $y_p = \sum_k w_{p,k} x_k = \boldsymbol{w}_p^T \boldsymbol{x}$ for a given set of parameters $\boldsymbol{w}_p$. Therefore, the first goal is to maximise the variance $Var(y_1)$ with respect to parameter $\boldsymbol{w}_1$:

$$\begin{aligned} Var(y_1) = \mathbf{E}\{y_1^2\} - \mathbf{E}\{y_1\}^2 = \mathbf{E}\{(\boldsymbol{w}_1^T \boldsymbol{x})^2\} = \\ = \boldsymbol{w}_1^T \mathbf{E}\{\boldsymbol{x}\boldsymbol{x}^T\}\boldsymbol{w}_1 = \boldsymbol{w}_1^T C_{\boldsymbol{x}} \boldsymbol{w}_1 \\ \text{such that } \|\boldsymbol{w}_1\| = 1 \end{aligned} \tag{2.1.1}$$

where $C_{\boldsymbol{x}}$ is the covariance matrix of $\boldsymbol{x}$. This is an optimisation problem which yields to the following Lagrange multiplier M:

$$M = \boldsymbol{w}_1^T C_{\boldsymbol{x}} \boldsymbol{w}_1 - \lambda(\boldsymbol{w}_1^T \boldsymbol{w}_1 - 1) \tag{2.1.2}$$

$$\frac{\partial M}{\partial \boldsymbol{w}_1} = 2C_{\boldsymbol{x}} \boldsymbol{w}_1 - 2\lambda \boldsymbol{w}_1 = 2(C_{\boldsymbol{x}} - \lambda I)\boldsymbol{w}_1 = 0 \tag{2.1.3}$$

Thus, the desired parameters vector is $\boldsymbol{w}_1 = \mathbf{e}_1$, i.e. the eigenvector of $C_{\boldsymbol{x}}$ which corresponds to the highest eigenvalue [16]. Moreover, it is desired that $\mathbf{E}\{y_p y_1\} = 0$ for $p > 1$. Therefore, for $p = 2$,

$$\mathbf{E}\{y_2 y_1\} = \mathbf{E}\{(\boldsymbol{w}_2^T \boldsymbol{x})(\boldsymbol{w}_1^T \boldsymbol{x})\} = \boldsymbol{w}_2^T C_{\boldsymbol{x}} \boldsymbol{w}_1 = \boldsymbol{w}_2^T C_{\boldsymbol{x}} \mathbf{e}_1 = \tag{2.1.4}$$

$$= \lambda_1 \boldsymbol{w}_2^T \mathbf{e}_1 = 0 \tag{2.1.5}$$

Thus, the second goal is to maximise $Var(y_2)$ looking into the subspace of vectors which are orthogonal to the first eigenvector of $C_{\boldsymbol{x}}$. The solution is, again, given by $\boldsymbol{w}_2 = \mathbf{e}_2$. By recursively applying these arguments for $p = 1, \ldots, n$, it follows that $\boldsymbol{w}_p = \mathbf{e}_p$, for all $p = 1, \ldots, n$.

Moreover, due to the recursive PCA definition, new variables in $\boldsymbol{y}$ are ordered depending on the data variance. The first variable (first principal component) is responsible for the major data variability. The second variable (second principal component) is responsible for the second major data vari-

ability, and so on. Therefore, variables in $\boldsymbol{y}$ can be neglected by thresholding the corresponding variability or according to the ordered cumulative carried data variability. Often, in practical cases, it turns out that a considerable number of new variables can be neglected, since the first few principal components carry, cumulatively, more than 95% of the whole data variability. This is particularly useful as dimensionality reduction method to pre-process data for machine learning, since the new set of transformed data carry almost the same data variability as the original set of data, with a smaller number of (uncorrelated) variables. Since it is not required to exploit labelled training data, PCA can be considered as an Unsupervised (U) method.

2.1.2 L_2-Regularised Logistic Regression

The logistic regression [17] is a statical method which establishes a mapping between input data, i.e. vectors $\boldsymbol{x}_i$, to a categorical label $l_i \in \mathcal{L}$, where $\mathcal{L}$ is a set of predefined labels. Therefore, logistic regression is mainly exploited in this thesis as a supervised classification method. The method aims to maximise a likelihood metric $L(\boldsymbol{w})$ with respect of parameters $\boldsymbol{w}$ as follows:

$$\max_{\boldsymbol{w}} L(\boldsymbol{w}) = \max_{\boldsymbol{w}} \log \prod_{i=1}^{n} \mathbf{P}(L = l_i | \boldsymbol{x}_i, \boldsymbol{w}) \tag{2.1.6}$$

$$\mathbf{P}(L = l_i | \boldsymbol{x}_i, \boldsymbol{w}) = \text{Logit}(\text{S}(\boldsymbol{x}_i)) = \frac{1}{1 + e^{-\text{S}(\boldsymbol{x}_i)}} \tag{2.1.7}$$

$$\text{S}(\boldsymbol{x}_i) = \boldsymbol{w}^T \boldsymbol{x}_i \tag{2.1.8}$$

where L is the predicted class for the sample $\boldsymbol{x}_i$ and S is the *score function*. Since no closed-form solution is available for (2.1.6), gradient ascent is used to find the best set of parameters $\boldsymbol{w}^*$ which maximises the metric. The regularisation problem aims to limit the so-called *overfitting problem*, which arises when a given model perfectly maps training data but fails in mapping testing data to the expected label. In other words, the model fits the training

set, but it is not able to extract general rules which can be potentially applied to correctly classify new data. To achieve this regularisation, it is common to consider the magnitude (L_2-norm) of the parameters as regulariser, i.e.

$$\max_{w} \left(L(\boldsymbol{w}) - \gamma \left\| \boldsymbol{w} \right\|_2^2 \right) \tag{2.1.9}$$

where $\gamma \in (0, \infty)$ emphasises how much the regularisation term best fits the problem. In particular, the regularising term penalises large coefficients, since the regulariser is subtracted from the likelihood metric. Typically, γ is chosen by using a validation set (for large datasets) or *cross-validation* (for small datasets). Since (2.1.7) uses the Logit function, binary classification can be performed with the model defined above. However, generalisations to multi-class classifications are based on, for example, Softmax and Cross-Entropy [13] functions. Overall, L_2-Regularised Logistic Regression can provide strongly non-linear mappings between input data and class labels, which makes it one of the first choices for supervised ML.

2.1.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) [17] is acknowledged as one of the most performing methods for supervised classifications [18]. It is based on the simple idea of maximising the geometrical margin between two classes in a simple binary classification problem for linearly separable data. In this case, SVM aims to find those two parallel hyperplanes (normal to the vector of parameters $\|\boldsymbol{w}\|$) such that their mutual distance is the highest and data points lie in the right side of them. In formulas, given a set of data point $\boldsymbol{x}_i$ and binary categorical labels $l_i = \{1, -1\}$, the aim is to find two parallel hyperplanes $\boldsymbol{w}^T\boldsymbol{x} - b = 1$ and $\boldsymbol{w}^T\boldsymbol{x} - b = -1$ such that their distance (which turns out to be equal to $\frac{2}{\|\boldsymbol{w}\|}$) is maximised. Therefore, with some algebra,

the problem can be formulated as

$$\min\|\boldsymbol{w}\| \quad \text{subject to} \quad l_i(\boldsymbol{w}\boldsymbol{x}_i - b) \geq 1, \ \forall i \tag{2.1.10}$$

The values $\boldsymbol{w}^*$ which solves (2.1.10) defines an hard-margin classifier such that

$$\boldsymbol{x}_i \leftarrow \text{sgn}(\boldsymbol{w}\boldsymbol{x}_i - b) \tag{2.1.11}$$

It can be shown that, in order to obtain a soft-margin classifier for non-linearly separable data, the problem can be re-written as

$$\min \left(\underbrace{\frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - l_i(\boldsymbol{w}\boldsymbol{x}_i - b)\}}_{\text{A}} + \underbrace{\gamma\|\boldsymbol{w}\|^2}_{\text{B}} \right) \tag{2.1.12}$$

where the term A ensures that data points lie in the right side of the decision boundary $\boldsymbol{w}\boldsymbol{x} - b = 0$, while the term B ensures that the size of the margin increases as much as required to reach an optimal trade-off. This method can also be generalised for multi-class problems and for non-linear supervised classification by replacing the dot product in the above-mentioned equations with *kernel functions* [13]. In this thesis, SVM is only exploited for semi-supervised anomaly detection. Therefore, non-linear binary classification is considered, with training data which includes *normal* instances only. The idea behind this so-called *one-class* SVM is to a priori define a percentage p of training data that is expected to be an outlier. During the training, the algorithm tries to train the bias term b in (2.1.10) such that p observations in training data have a negative score. Such standard method is common for semi-supervised anomaly detection [14], and it has been widely exploited in Chapter 5.

2.1.4 Deep Neural Networks

Deep Neural Networks are widely studied and exploited in several fields nowadays [13]. A Deep Network is made by connecting several layers of interconnected processing units called *neurons*. A neuron receives numerical inputs from previous neurons and then processes it by using linear and non-linear operations, before sending it to the next layer with which it is connected. A deep network which includes a sufficient number of neurons and layers is, in principle, able to map input data to output data of any sort, e.g. labels in the classification case, replicating a wide-range of non-linear functions. However, the advantage of Deep Networks is that it is possible to *automatically* and *implicitly* establish such complex mappings by using *backpropagation* algorithm [13]. In short, given a certain loss function which measures the difference between the expected output, e.g. a pre-defined classification label, and the actual network output, e.g. the predicted label, it is possible to repeatedly adjust internal network parameters in order to minimise the loss function, i.e. minimise the errors between expected and predicted output. At each iteration, the network's performance is measured, e.g. by testing it on a pre-defined validation dataset. Thus, when a trained network reaches a validation performance which is acceptable for the given application, the iterative process is stopped, and a conclusive performance evaluation is conducted on a pre-defined testing dataset. As a result, this iterative process requires consistent amount of data due to: 1) separated training, validation and testing data are required; 2) if testing data is well defined including as general instances as possible, high testing performance is the result of high network's generalisation abilities, which in turn are most likely to be obtained by consistently increasing training and validation data samples. Overall, the network internal parameters definition is performed implicitly by showing several input examples and expected outputs to the network during the training phase. In Fig. 2.2-(a), an example of a simple,

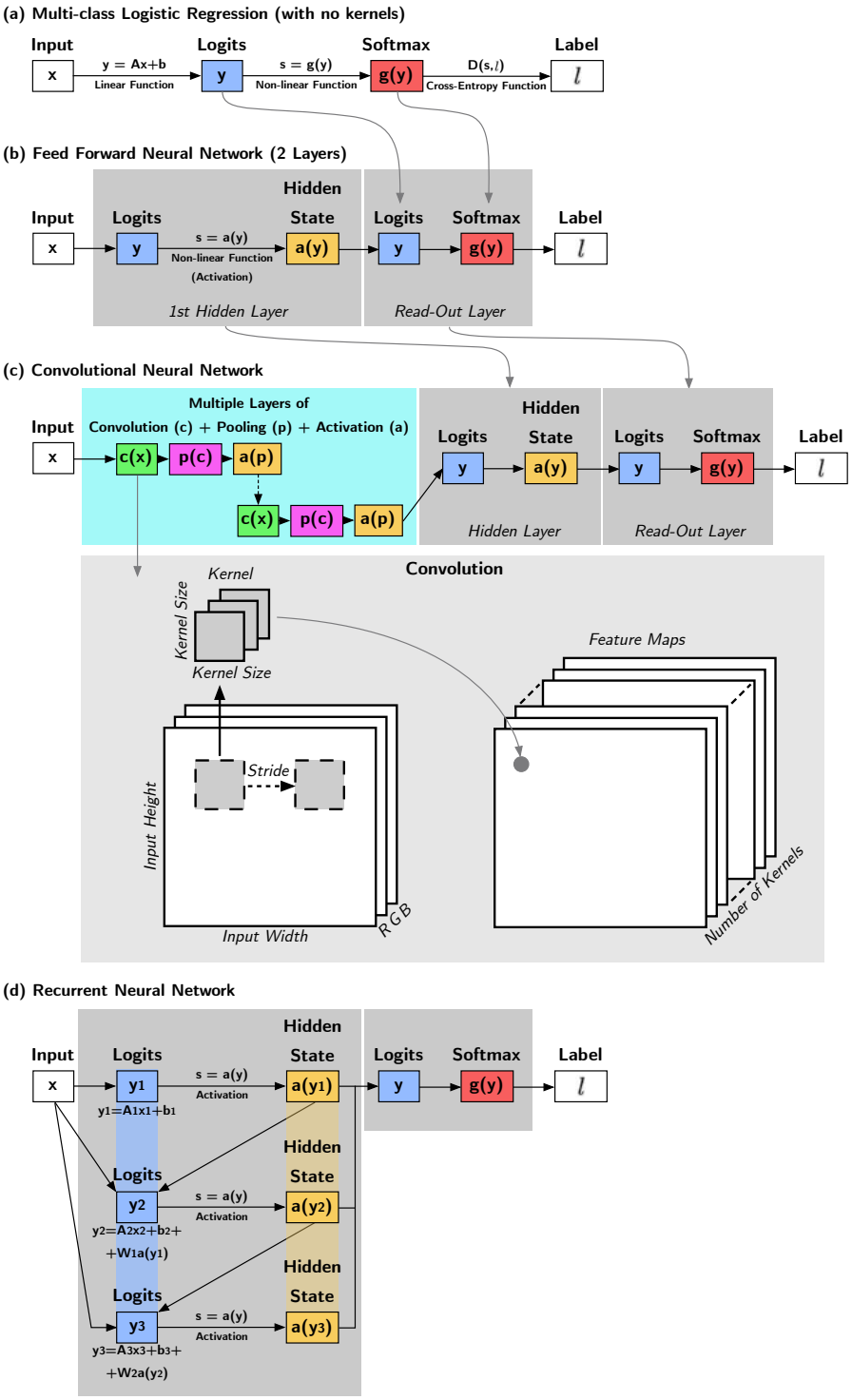


Figure 2.2: (a) Multi-class Logistic Regression via neural network layers, with no kernels. (b) Example of Feed Forward Neural Network (FFNN) with one hidden layer and a read-out layer, which in turn is made by using fundamental blocks for Logistic Regression. (c) Example of Convolutional Neural Network, where an FFNN is following the convolutional and pooling layers. The convolutional layer is further expanded to see the basic convolutional operations based on kernels. (d) Example of basic Recurrent Neural Network, where the input $\boldsymbol{x}$ is a time-sequence $\boldsymbol{x} = [x_1, x_2, x_3]$.

logistic regression implemented via neural networks for multi-class classification is shown. This processing core is normally exploited as a read-out layer in any neural network. For example, in Fig. 2.2-(b), a simple feed forward neural network is depicted, which uses the above-mentioned core after a single hidden processing layer.

Remarkable examples of Deep Networks are the Convolutional Neural Networks, which are specifically designed to efficiently deal with images as input. As shown in Fig. 2.2-(c), the input is given to the first network layers to performs multiple convolutions and pooling operations, followed by non-linear steps in which *activation* function introduce non-linearity. In this case, typically, the convolutional layers are followed by a feed forward-like structure, to transform feature maps (tensors) provided by the convolutional layers into class labels (vectors).

Further evolution of the feed forward neural network is provided by the Recurrent Neural Network (RNN), which are a particular class of Deep Networks which are designed to deal with temporal input sequences, e.g. time-steps based input data. As shown in Fig. 2.2-(d), the basic idea is that, at each time-step, the output of the recurrent layer is provided as input again to the same layer, alongside the current time-step sequence element. Therefore, the output of the previous time-step acts as a *memory* and it affects the computation of the current time-step. It turned out that the major drawback for RNNs was the vanishing gradient [13]. In short, the conventional RNN suffers from losing the information from the previous time-steps, i.e. the memory, pretty quickly. Therefore, the Long Short-Term Neural Network (LSTM) was designed to overcome the problems occurring in the classic RNN. This network has the advantage to allow more flexibility in the way the memory from the previous time-steps is managed. This advance yields to a network which is able to keep in memory information for a long or short time, avoiding memory corruption.

Interestingly, combinations of CNNs and LSTMs have been exploited for video-based classification, since a video is naturally defined as a time sequence of subsequent images. Therefore, the CNN structure can be exploited as a feature extraction method on the spatial domain (frames), and the LSTM can be exploited for modelling temporal dependencies between extracted spatial features. However, several other structures have been studies, achieving different performance depending on the domain [19].

2.1.5 Autoencoders

The autoencoders are special cases of Deep Networks, particularly useful for dimensionality reduction, de-noising and anomaly detection [13]. The main concept is based on jointly training a neural network-based *encoder* and a *decoder*, with the goal of obtaining an output which is as close as possible to the correspondent input. The idea is that the encoder maps input data to a more compact representation, which is expected to contain only information which is relevant to the task. As opposite, the purpose of the decoder is to reconstruct the original input starting from the compact representation provided by the encoder. By training both encoder and decoder simultaneously, a trade-off between encoding and decoding performance is achieved, and a tool to provide compact transformations of data is provided.

Autoencoders are used in this thesis as baselines for performance comparisons in Chapter 5 regarding HAAD. In fact, CNN-based autoencoders are acknowledged as effective tools for video-based semi-supervised anomaly detection. The idea is that, once the autoencoder is trained on normal data, new data which resembles training data are expected to be reconstructed with low error. In contrast, outliers are expected to be reconstructed with higher error, since the network is not optimised to reconstruct such data. As discussed in Chapter 5, building upon this idea, numerous approaches have been designed, with a specific focus on video-based anomaly detection.

2.1.6 Hierarchical Clustering (HC)

Hierarchical Clustering (HC) is a simple and well-known algorithm that aims to explore data points in a multidimensional space $\boldsymbol{x} \in \mathbb{R}^n$ looking for local groups, i.e. clusters. HC is based on an exhaustive search within data based on mutual distance, e.g. Euclidean distance. The algorithm starts considering all data points as separated clusters. Therefore, it finds the two closest clusters, merging them into a single cluster. Thus, it repeats this operation until only one cluster is defined, containing all data points. Typically, HC output consists of a *dendrogram*, i.e. a tree diagram which represents the mutual connections established by the algorithm. Therefore, by setting an initial parameter p as the number of data clusters are expected in the dataset, the dendrogram can be used to find the unique cluster configuration which corresponds to the requested number of clusters p (Fig. 2.3). HC per-

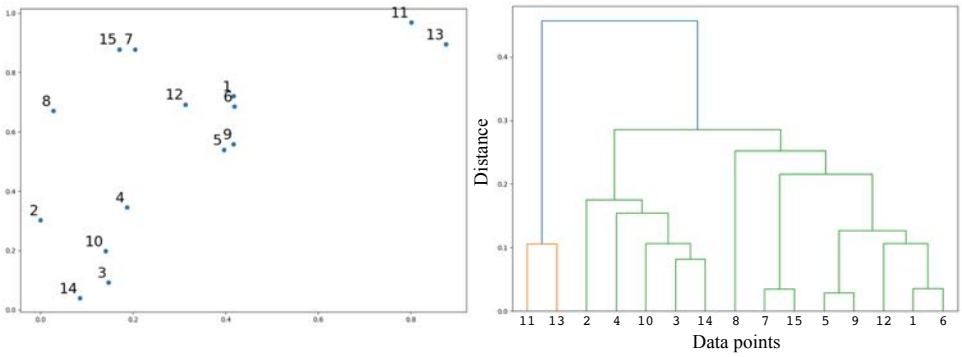


Figure 2.3: (Left) Example of 2-dimensional data points, numbered from 1 to 15 for illustration purpose. (Right) Correspondent dendrogram based on mutual distances between data points. In this example, $p = 2$; thus, 2 clusters are identified by using the dendrogram, i.e. yellow and green dendrogram leafs.

forms unsupervised clustering, since no internal descriptions of data points, i.e. labels, are required to set the clusters. In contrast, clusters are defined on the basis of the mutual similarity in terms of distance. However, prior knowledge is required or multiple thresholding parameters p are required to find the best number of expected clusters [20].

2.1.7 Self-Organising Maps (SOM)

The Self-Organising Maps (SOM) is a popular artificial neural network-based unsupervised features selection method for clustering and dimensionality reduction. The goal is to find optimum mappings between high-dimensional input data points and a low-dimensional predefined discrete representation based on graph-like *nodes* [21]. SOM is based on competitive learning, in contrast with back-propagation with gradient descent based optimisation commonly used for training CNN and RNN networks.

The basic idea behind SOM is to mesh a predefined uniform lattice on top of a multidimensional data point distribution. In practical implementation, SOM requires to input the predefined lattice, which is expected to be an overestimated set of clusters. The lattice is normally a two-dimensional, regular, rectangular or hexagonal. However, in principle, lattices of any dimension and shape are allowed. Once the mapping is produced, each data points can be mapped to one of the lattice nodes, which effectively represents a first clustering abstraction level. Subsequently, HC or other methods such as K-means can be further applied to further reduce the number of clusters and create the second clustering abstraction level. SOM can be particularly effective as dimensionality reduction for subsequently applying HC since HC tends to be particularly computational expensive when high-dimension and high-number of data points are considered [22].

2.2 Human Action Recognition and Activity Detection

2.2.1 Overview

Video-based Human Action Recognition is a vast field. It comprises several subdomains, depending on the source input, the main used techniques and the desired application. In this section, a general overview of the field is provided, with major attention on silhouette and pose based HAR, which is

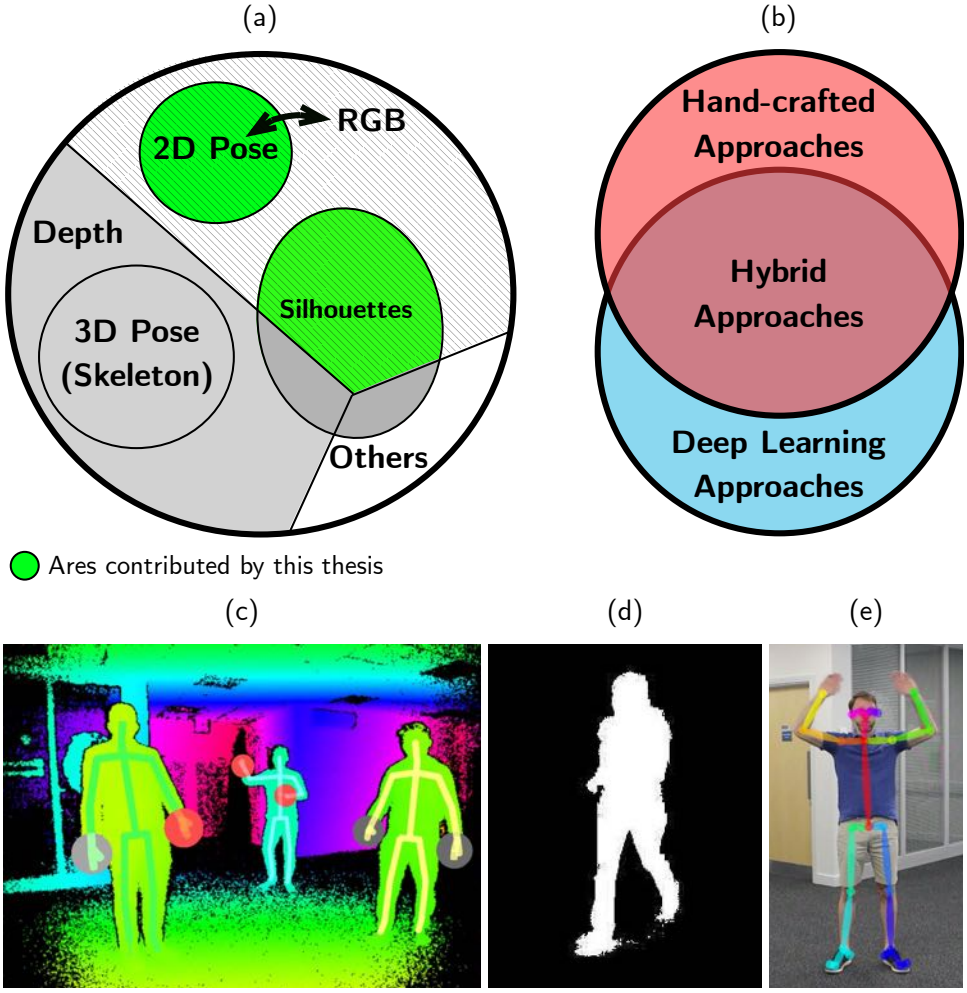


Figure 2.4: (a) HAR input data overview and areas contributed by this thesis. (b) HAR features extraction approaches overview. (c) Depth point cloud (Kinect) and estimated human skeletons, where human body silhouettes are visible. (d) Human silhouette detected from RGB data via background subtraction. (e) Human pose detected via RGB-based human pose detectors.

the focus of this thesis.

As depicted in Fig. 2.4-(a), the two major types of input data for HAR are conventional RGB videos and depth data [23]. Human action information can be stored as RGB-based data or as point clouds (3D data points) provided by depth sensors, e.g. Structured-light cameras such as Kinect (Fig. 2.4-(b)) or Time-of-flight cameras. The major difference between RGB and depth data is that the latter is a 3D spatial representation of the scene, including the human targets, while the RGB data are a 2D three-channels-

based representation. Therefore, in the RGB case, information about the depth is mostly lost. This has several negative impacts, for example: 1) actions that involve body movements *toward* the camera are more difficult to be taken into consideration; 2) the 3D volumetric body shape is lost; 3) distinctions between targets in partial occlusion each other are more challenging. Therefore, despite depth data advantages, RGB data is collected by a conventional camera instead of more expensive depth devices. Moreover, depth cameras working range is short, and they suffer from outdoor light negative effect [5], while RGB cameras can still provide useful information from a further distance and in a wider range of illumination conditions.

From the human perspective, raw RGB data represents a superb source of human action information, since the sight is the sense that humans typically use to gather information about the outside world. However, as a matter of fact, the human body is an extraordinarily complex and variable object, which we, as humans, require years of learning and training to master and understand. Therefore, from the artificial intelligence point of view, RGB data often comprises a multitude of cluttered and irrelevant information which hides human action information. For example, the first non-trivial task for an AI is to *detect* a human target, before trying to understand what the target is doing in terms of actions. Such detection is instead trivial for human beings. Moreover, light condition changes and occlusions might compromise the detection very easily, while, for humans, it is quite obvious to figure out which is the target and what is doing even in the presence of challenging conditions. Furthermore, appearance and body shape changes, attitude, viewpoint changes, spatio-temporal complex and unexpected movements might quickly compromise the recognition and classification if no specific solutions are adopted [23]. Last but not least, the quality of the data might further exacerbate the problem. In contrast, the human brain can easily exclude these disturbing factors and reach the kernel

of the desired information.

As a useful refinement of RGB data, the shape of the body can be extracted and exploited for HAR. In this sense, the body silhouette is defined as the semantic area/volume, which refers to the body of the human target. The human silhouettes have been investigated for years as input for HAR algorithms, since they convey important features regarding the type of performed action, without the disturbing effect of other irrelevant information. From a simple visual comparison between Fig. 2.4-(c) and Fig. 2.4-(d), it can be seen that human silhouettes can be effectively be retrieved from both depth and RGB data [24]. However, in this thesis, only silhouettes retrieved from RGB data will be considered. In Section 2.2.2, an overview of the silhouette-based state-of-the-art is provided.

Despite the above-mentioned depth data limitations, Fig. 2.4-(a) shows that depth data are generally the precursor of the 3D poses or *skeletons*, which is a compact representation of the human body as a list of 3D coordinates. Skeleton-based HAR is an important subdomain, which inspired this thesis contributions. A human skeleton representation is provided as a set of 3D coordinates, i.e. *landmarks*, which locate key points of the body, such as *head*, *neck*, *torso*, *left-hand* etc., within the 3D space. Until a few years ago, landmark-based HAR corresponded to skeleton-based HAR, because the only reliable device for collecting body landmarks was a depth camera. However, with the advent of deep learning-based human body limbs detectors such as OpenPose [12], it became also possible to collect body landmarks from RGB data, i.e. *2D poses*. Similarly to a skeleton, 2D poses are provided as a list of 2D coordinates identifying the body limbs position in the frame. Despite skeleton and pose are similar representations, they mainly differ from the fact that human poses do not allow 3D computation, e.g. rotations in the 3D space, therefore the camera viewpoint is fundamental; in contrast, skeleton are embedded within a 3D space which allows more

computation flexibility and a certain degree of viewpoint insensitivity. In Section 2.2.3, an overview of the pose-based state-of-the-art is provided.

Overall, once the desired input data has been selected, e.g. human silhouettes, skeleton or poses, it is crucial to extract meaningful features from it. Such features are numerical quantities which are expected to be highly informative regards the performed action. In other words, a complex mapping between input data and tensors (features) is required. Therefore, once the mapping is provided, features-based classification can be performed. The classification is generally as much easy to be performed as much the data-to-features mapping is informative. In other words, if dissimilar action samples are mapped to similar or identical features, then the classification is destined to fail. Thus, it is crucial to design mapping systems that are highly descriptive, to ensure the classification step success.

Narrowing on RGB-based HAR, state-of-the-art algorithms can be divided into three main categories (Fig. 2.4-(b)), depending on how such a mapping is defined:

- *Hand-crafted features extraction.* In this category, the mapping is strongly based on human insights regards HAR problems, while machine abilities are not fully exploited and are often limited to conventional machine learning tasks. In other words, the solutions included in this category are based on local representations, which tries to focus on input details which are expected to be highly informative from the human perspective [25]. It is the case of the approach suggested by Weinland [26] and discussed Section 2.2.2.

- *Deep-learning features extraction.* In this category, Convolutional Neural Networks (CNNs), generative models, 3D-CNNs and Recurrent Neural Networks (RNNs) [27] are used to explore data without any or with very limited human insights. Therefore, the input-to-features mapping

is efficiently and automatically created during the network training, in a way that might not be any longer interpretable by humans. Despite clear advantages in terms of performance, processing explainability becomes an issue [28].

- *Hybrid features extraction.* The algorithms in this category attempt to combine the most promising results from both the above mentioned hand-crafted and deep learning-based approaches, providing a useful trade-off between them [25]. This thesis contributions mostly belong to this category.

Despite the variety of available techniques mentioned above, generating a good mapping between input data and features is always extremely difficult. In particular, the challenges which often compromise the mapping effectiveness are:

1. *Human appearance changes.* The system is expected to be robust to body size, body shape, gender, ethnicity and other appearance changes; therefore, the mapping must take into account that the same action might be performed by different human targets with different body appearance;

2. *Intra-class variations.* Humans have different attitudes due to different level of ability or experience with the performed action, e.g. an athlete jumps or runs differently compared to a sedentary person; the mapping has to take into account this additional effect;

3. *Action timing, speed, contextual changes.* Different subjects can perform the same action with different timing or speed. Even contextual differences affect the performed action, resulting in a huge variety of action styles;

4. *Occlusions, self-occlusions, missing data.* Occlusions occur when the

human target is partially or completely covered by contextual objects. Similarly, self-occlusions occur when the human target partially occludes itself due to the assumed position with respect to the camera. Missing data occur when the human target is partially out from the camera field of view.

5. *Viewpoints changes.* Human actions might look extremely different from different camera viewpoints, resulting in a severe drop of information when the viewpoint is not the best possible; therefore, proposed solutions are required to take into account this variability;

6. *Camera moving and zooming.* Movements of the camera, including zooming, can easily drop down performance;

7. *Background clutter.* This issue occurs when the subject is not clearly distinguishable from the background or when the background is moving, generating false positive subjects/targets; in this case, input data might be noised or corrupted, compromising further processing;

8. *Generalisation.* Several trained models are scenario-specific or data-type specific. Thus, generalised models which can be tested over different scenarios/datasets are desirable for effective applications;

9. *Models explainability.* Deep learning methods, which are predominant in the current literature, have great performance at the expenses of understandability of the learning process itself. On the contrary, hand-crafted learning, which has been very popular until a few years ago, in general, can be considered less generalisable and more data-type specific. However, they are better understandable from the human side. What is the best trade-off approach is still an open question;

10. *Time localisation.* In a continuous data stream where targets might perform multiple actions, it is crucial to identify the extent of the

action, i.e. *when* it starts and *when* it ends.

Despite the efforts to address these challenges, no definitive solution has been presented to address HAR. However, extensive studies suggests that *multimodal data* improves performance [25]. For this reason, this thesis focuses on extracting multi-semantic data from RGB data, e.g. human silhouettes, human poses, and relevant objects bounding boxes, to be processed alongside RGB data itself in a multimodal approach, to improve HAR and HAAD performance.

With the recent advances of Deep Learning, research in HAR started exploring these techniques for digging into raw RGB data, leveraging the superb abstraction power of neural networks. In a recent and famous advance from Carreira et al. [19], existing deep learning techniques are reviewed, and new advances are presented. These methods work in wild and cluttered scenarios. RGB-based deep learning architectures are based on combinations of Convolutional Neural Networks (CNNs), 3D-Convolutional Neural Networks (3DCNNs) and Long Short-Term Memory neural networks (LSTMs), often exploiting pre-training as beneficial warming up method [29]. These networks are trained on large datasets, such as Kinetics [30] and UCF101 [31], containing hundreds of complex actions with a consistent amount of contextual information, with different camera proximity and a variable number of targets. Despite such challenging scenarios, the performed HAR practically consists of video clip *captioning* [28]. Therefore, the trained model can, in principle, detect if one of the trained action has been performed in the video, providing additional descriptions about the context. However, using the whole frame as a unimodal source of information does not provide hints on how the system can discriminate between different human targets actions. In contrast, if human detections are available and human tracking is performed, as in surveilled environments, in principle the bounding-boxes RGB data can be exploited as input instead of the whole frame [32]. How-

ever, this thesis provides evidence that this approach is suboptimal to detect human posture-related actions, compared to the pose-based approach. RGB data might be a too complex and cluttered source of information. Therefore, posture-related information extracted from it, such as human silhouettes and human poses, might be a better source of information, since disturbing factors are mostly neglected. However, RGB data can still carry additional knowledge that can be potentially be further exploited to compensate what silhouette and poses might involuntarily neglect. Therefore, hybrid solutions between existing hand-crafted techniques and recent deep learning advances can be leveraged, to reach those levels of multimodal understanding which is desired, for example, in surveillance scenarios. For this reason, one of the final goals of this thesis is to explore raw RGB data combined with pose-based data for posture-related HAR, to take the most from both modalities in a multi-target monitored environment.

The importance of human postures is well-known in literature and is deeply rooted in other fields of computer vision. In [33] the authors used gesture cues to recognise who is talking in a scene. However, using gestures and postures, we can reveal more complicated social behaviours and even attitudes, even recognising an extrovert or introvert individual [34]. Posture and gestures are social cues which are able to carry information about how we relate to the environment and how we feel in a particular situation [35]. In [36] the authors dealt with gestures and postures to determine if a group of people is interacting with each other, also considering when the visual focus of attention for different individuals are intersecting, people are considered in interaction. In [37, 38] studies related to groups of interacting people are explained, exploiting their attitude in occupying the common 3D space: it turns out that interacting people are recognisable through the shape of the area between them and how they stand in that area, i.e. the mutual *silhouette*.

In the following sub-sections, the silhouette-based and pose-based HAR state-of-the-art is further discussed.

2.2.2 Silhouette-based HAR

Silhouettes have been repeatedly exploited for HAR, since they carry a considerable amount of information related to the body shape, and therefore, to the posture-related action performed by the target. Generally, human silhouettes can be extracted by using background subtraction algorithms, such as the patented ViBe [39]. Modern background subtraction methods generally distinguish foreground from background as a pixel-based classification problem. However, as already discussed in the previous section, depth data can also be the precursor of human silhouettes. Moreover, Batchuluun et al. [40], demonstrated that human silhouettes could also be effectively retrieved from infrared and thermal cameras for effective HAR in surveillance environments.

In a multimodal-data based interesting work, Tang et al. [41,42] proposed a borrowing information method to share knowledge across modalities. In particular, depth-skeleton information stored in an offline action database was leveraged to improve the accuracy of RGB based HAR. The idea was to exploit depth and skeleton data stored in the offline dataset to support the RGB based processing. Despite the multiple sparkling points behind this work, it is crucial to highlight that the RGB based HAR was performed by using human silhouettes, following the well-established framework previously provided by Weinland et al. [26, 43, 44]. Tang et al. effort was motivated by the fact that, despite Weinland et al.'s method was effective, it provided consistent room for improvements. The Weinland et al. method was one of the most interesting outcomes for 2D RGB based HAR before the advent of deep learning as a trending technique for HAR [45]. It was specifically designed to be robust to occlusions and viewpoint changes. The core idea was

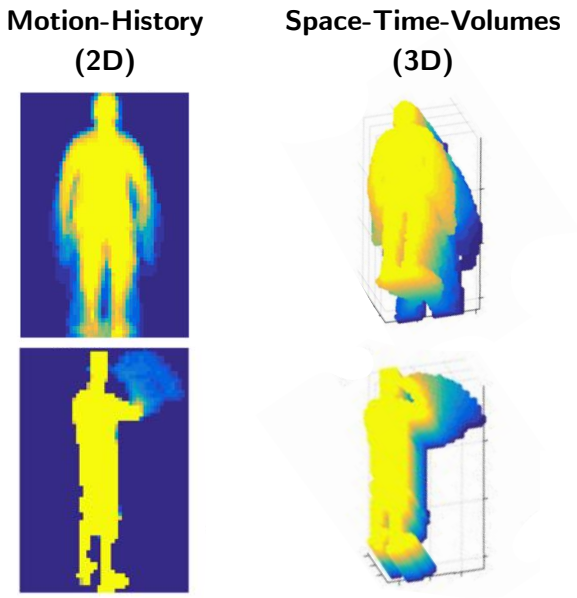


Figure 2.5: (Left) Motion-History maps obtained by piling up human silhouettes at different time steps and averaging on the temporal dimension. (Right) Space-Time Volumes (STV) obtained by piling up human silhouette at different time steps.

to implement local 3D Histogram of Oriented Gradients features (3D-HOG) deployed in a framework with multiple classification models, depending on the body location. Despite the claimed robustness of this approach, my implementation of this system highlighted limitations in attributing adaptive weights for the local gesture contribution to the overall decision. In Chapter 3, these limitations are discussed, and alternative solutions are proposed.

3D-HOG features are the 3D extension of their 2D counterpart, Histogram of Oriented Gradient (HOG) [46], developed by Klaser et al. [47]. The original version of HOG as been widely used in literature for 2D black-white image features extraction [48–51]. In all these works, the action is represented as a *Motion-History* maps (Fig. 2.5-(Left)), which encodes the temporal changes of the body silhouettes with colour shades to encode the movements. Therefore, the main idea of HOG features is to encode local colour shades changes by projecting local gradients onto pre-defined circular directions, and defining histograms which cumulates local projection magnitudes.

As opposite, 3D-HOG features, can be used to extract features from target *Space-Time Volumes* (STV) [52] (Fig. 2.5-(Right)). Silhouette-based STV is a 3D human action representation made by piling up target silhouettes at consecutive frames. Therefore, the resulting STV represents a 3D data volume, and 3D-HOG features can be computed to extract features related to the STV local shape. However, 3D-HOG have also been used in literature for different computer vision problems. For example, an object detector has been developed based on 3D-HOG features applied to depth volume voxels [53]. An evolution of this method was presented by Dupre et al. [54], who designed a method for risk estimation in the presence of sharp objects. Furthermore, since depth-based action data are represented as a 4D-tensor, where the fourth dimension represents the time, straightforward generalisations of the 3D-HOG theory are possible. For example, a 3D-HOG generalisation to deal with the fourth dimension was proposed by Oreifej et al., who proposed the so-called HON4D [55].

One of the advantages of 3D-HOG features for HAR is related to the robustness to multi-viewpoint actions [26]. In general, multi-viewpoint action recognition relies on finding common features among videos from a different point of views [56]. However, 3D-HOG features can be exploited to extract peculiar features from each viewpoint data and to train a general classifier on multi-viewpoint extracted features, as described in Chapter 3. In contrast, Iosifidis et al. [57], exploited human silhouette to propose a multi-viewpoint framework which separately learns from each viewpoint data. Therefore, to classify a new sample recorded from a specific viewpoint, a voting strategy between multiple models was required. Differently, Azary et al. [58], proposed a sparse representation which was able to train a single model with samples from for up to 8 viewpoints, obtaining competitive accuracies. The approach proposed in Chapter 3 follows the same idea but obtains outperforming results. In very recent work, Chou et al. [59], developed a hybrid

approach considering again human silhouettes alongside points of interest within RGB data to train a multi-viewpoint model for HAR. Points of interest are another powerful, view-invariant features which can be useful to provide RGB-based, multi-viewpoint models. However, the points of interest are beyond the scope of this thesis.

Another notable approach exploiting human silhouettes was provided by Chaaraoui et al. [60] who proposed a coevolutionary algorithm for HAR. In this work, bags of silhouette were defined for each action classes, in order to select a few key silhouettes out from training data and reduce the overfitting problem. This technique was also at the core of the Weinland et al. approach, and it is further exploited in the proposed technique in Chapter 3.

Recently, Jin et al. [61], exploited deep learning to deal with silhouette based sub-action HAR. Jin et al. highlighted that human actions could be actually seen as combinations of stratified effects due to the posture, the locomotion and the semantic gesture levels. Therefore, Jin et al. proposed a framework to assign multiple labels to each testing action, defining a sort of complex caption of the target actions. Despite the effectiveness of this work, it was not possible to further explore this interesting idea due to lack of multi-labelled data. However, Jin et al. system was based on human tracking, to allow multiple target HAR in CCTV environments. This approach links with the solutions discussed in Chapters 4 and 5, since similar frameworks are exploited to deal with multiple target data.

2.2.3 Pose-based HAR

Pose-based HAR is strongly related to skeleton-based HAR, which in turn is related to depth-based HAR. In the recent past, several contributions have been made by depth-based HAR community [62–64]. However, despite the significant advantages provided by the 3D depth data and human skeletons, such as viewpoints robustness and generalisation, insensitivity to light con-

dition changes and the clear benefits provided by the third dimension [62], depth devices, e.g. Kinect, have critical drawbacks. For example, Kinect does not work well in outdoor environments and has a minimal working range (up to 5-6 meters), which limits its ability to be implemented in surveillance scenarios [5]. Motivated by the limitations of depth-based devices, researchers started developing new techniques that can provide 2D poses from RGB data. Poses are qualitatively similar to skeleton data, although the third dimension is lost. Therefore, highly promising body pose detectors have been published in recent years, such as DeeperCut [65] and OpenPose [12]. In particular, OpenPose achieves the best performance, opening new research directions for HAR. As a matter of fact, human poses represent an effective way to extract posture-related information from RGB data, exactly as 3D skeleton do from depth data (Fig. 2.6). However, it is worth emphasising that 2D poses have limitations compared to skeletons. For example, 2D poses do not allow any rotation in the 3D space, which in contrast are allowed for skeleton data [67]. Thus, many skeleton-based algorithms, which are explicitly or implicitly based on the advantageous properties of the 3D space, cannot be directly applied to the 2D case. Furthermore, no depth data, i.e. Kinect points cloud, is available in the 2D case. Therefore, depth-based feature extraction theory cannot be directly applied to the RGB-based case. Despite these problems, some of the fundamental skeleton processing can still be lent to pose-based HAR. For example, landmarks normalisation is commonly used to remove target size and location dependency from skeleton data, and it can be easily replicated in the pose-based case. Wang et al. [68] proposed to extract Local Occupancy Pattern (LOP) features from depth data and Invariant Features from skeleton data and depth data, where skeleton normalisation was used. Similarly, Taha et al. [69] suggested implementing Hidden Markov Models (HMMs) to process skeleton data where normalisation was performed as a pre-processing step.

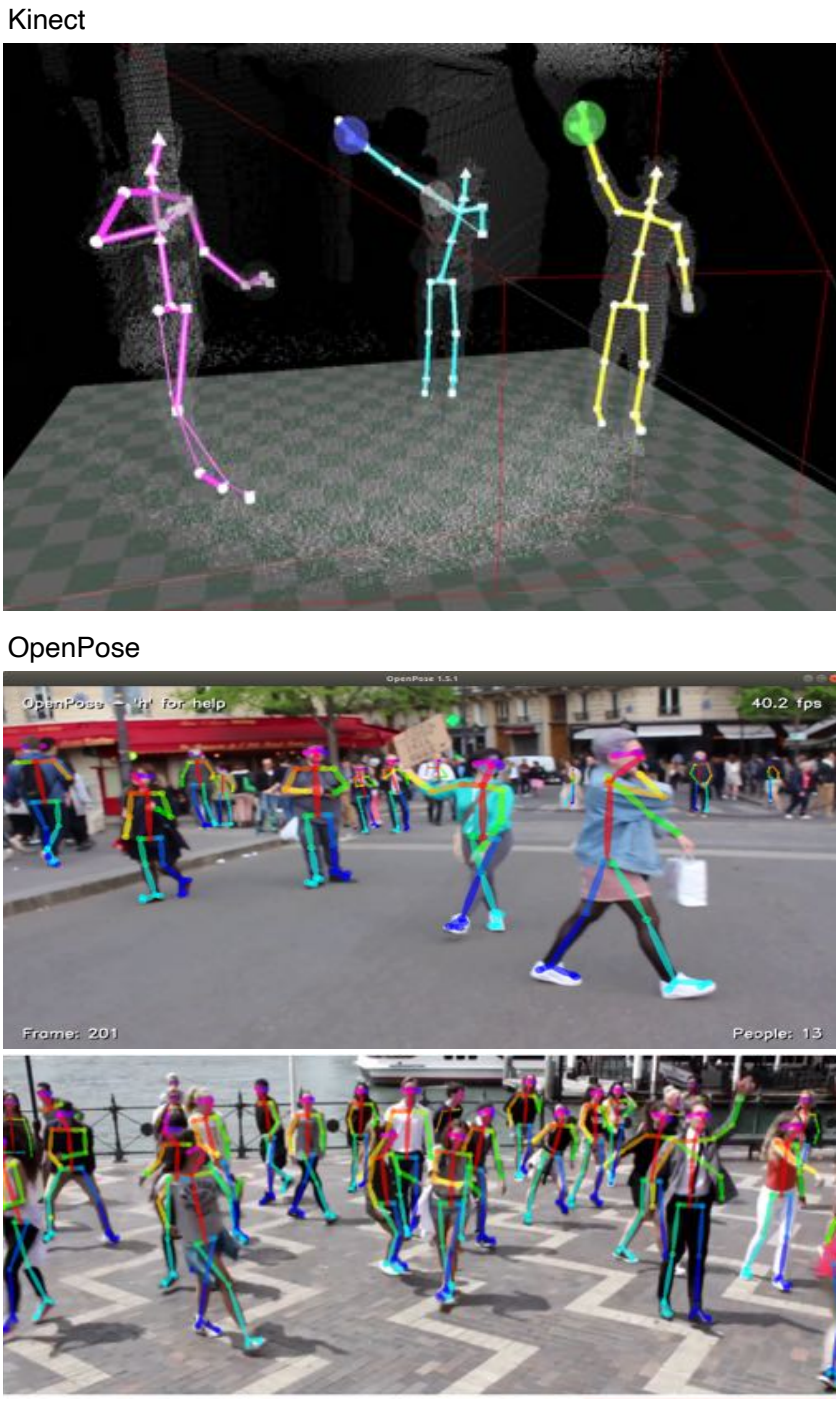


Figure 2.6: (Top) Kinect data (point cloud) and estimated 3D skeletons for three human targets. (Bottom) 2D human poses estimated from RGB data by OpenPose [12, 66].

The work presented in this thesis regarded pose-based HAR is among the few more other parallel attempts of exploring this new research opportun-

ity. For example, Yan et al. [70] have implemented graph convolutional networks to process pose data provided by OpenPose, which achieves promising results. This approach considers temporal information alongside spatial information embedded in a common structure. However, this method requires a fixed number of frames for each action sample in order to build the action graph, which may affect system flexibility. In contrast, the proposed solution in Chapters 4 and 5 focuses on exploiting LSTM based networks to deal with time-based sequences. This approach is specifically designed to deal with multivariate temporal sequences, with no restrictions on the number of time steps, allowing full flexibility with respect to space and time.

Regarding the classification step, RNN and LSTM have already been exploited in many recent papers related to skeleton-based HAR. Liu et al. [71] proposed the 3D skeleton-based action recognition using LSTM networks. Their major focus was on implementing trusting gates on the LSTM architecture to allow better action representation. Yong et al. [72] focused on developing a hierarchical RNN approach to focus on several 3D skeleton subparts, in order to better discriminate which body sub-part is related to the performed action. Veeriah et al. [73], authors focused on new gate strategies for LSTM, in order to emphasise salient motion from learning data. Zhu et al. [74], new regularisation term for LSTM was used in order to learn co-occurrences within skeleton data. Most of these above-mentioned studies focused on modifying the LSTM network architectures to process the raw 3D skeleton data more effectively. In contrast, in Chapter 4, the proposed work aims to extract more meaningful and robust input features based on the 2D body landmark coordinates, to be provided as input for a pre-defined LSTM architecture.

Other approaches relevant to the Chapter 4 for performance comparison over the same datasets are [75–83]. In these works, the learning sources was mainly raw RGB data, motion history maps and body silhouettes. Compared

with human poses, these data are generally heavier, redundant and affected by noise. In particular, despite data redundancy can be advantageous for HAR, more processing time is required. In this sense, 2D pose data are extremely efficient to be processed for classification. Moreover, in the above-mentioned works, no explicit link with human tracking was given. In Section 2.2.5, a brief overview of multiple human tracking is provided. In contrast to the above-mentioned baselines, the approach proposed in Chapter 4 is specifically designed for tracking-based frameworks, allowing straightforward implementation in multi-target scenarios.

2.2.4 RGB vs Silhouette vs 2D Pose

On the basis of the arguments discussed in Sections 2.2.1, 2.2.2 and 2.2.3, a direct comparison between the mentioned modalities, i.e. RGB, human silhouettes and 2D poses, is provided. In particular, in this section, the rationale followed during this thesis exploration is summarised. Since RGB data are assumed to contain the whole available information, it is expected that a model trained on RGB can potentially reach the best performance for posture-based HAR. However, as this thesis demonstrates in Chapter 5, RGB-based posture-based feature extraction mostly end up being a deep learning-based body shapes/edges/silhouette detector, with the addition of a classification layer. Therefore, it is reasonable to argue that performing an explicit body silhouette detection, e.g. background subtraction, is convenient since the system is explicitly directed towards posture-related important data. Explored silhouette based feature extraction, i.e. 3D-HOG computed on STVs, was practically looking at body shapes changes. Such changes are in turn due to body limbs positions changes in the space-time domain. The experienced drawback of this approach was related to the background subtraction method, which was unreliable in practical scenarios. Therefore, a replacement for body silhouette was required, which was expected to solve

background subtraction limitations and still able to carry information about the body limbs movements. Therefore, since 2D pose is able to track body limbs position changes from RGB data, it was natural to argue that 2D pose might have been a good replacement of body silhouettes. In fact, this thesis proves that 2D poses based classification can obtain similar or outperforming performance compared to body silhouettes, yet providing a direct solution to the background subtraction limitations. Last but not least, by using RGB-based detectors, it was possible to extract poses from RGB data rather than from depth data as commonly performed by other state-of-the-art methods. Furthermore, once a robust method for pose-based HAR was defined (Chapter 4), it was reasonable to argue that RGB data might be further exploited to compensate information which is missing in 2D poses, e.g. finer details, appearance changes, and contextual information. Therefore, this thesis exploration closed the path where it began, by combining RGB and pose data, for combined HAR and HAAD.

2.2.5 Human Poses and Multiple Target Tracking

As already mentioned in the previous sections, Chapters 4 and 5 contributions rely on popular pose detectors such as OpenPose [12], to extract human pose data from RGB videos. Similarly, the popular object detector SSD [10] is an object detector which provides bounding boxes, i.e. Region Of Interest (ROI), and labels for a wide range of objects. SSD is exploited in Chapter 5 to extract the key object position from RGB data, to be fed into the learning process for object position anomaly detection.

The above-mentioned detectors provide frame-by-frame data for a variety of human targets and objects. However, humans and objects need to be tracked frame-by-frame [84–86], to obtain consistent target's identities. This step makes sure that data from two or more similar targets, e.g. two chairs or multiple humans, are correctly grouped over time, on the basis of the

tracked identity.

Although multiple human tracking is not the core of the proposed work, it is fundamental to allow preliminary steps when multiple human poses are detected in the scene. To the purpose of this work, it is worth mentioning a simple and effective tracker proposed by Bewley [87], based on Kalman's Filter and Hungarian algorithm. This algorithm will serve the proposed work, providing real-time and computationally light tracking.

Overall, it is worth mentioning that human poses are a good source of information to perform multiple human tracking [84, 85]. As further discussed in Chapter 5, tracking is a required pre-processing step for poses. Tracking allows long-time pose-based HAR since it is able to provide consistent targets identities and correctly assign multiple human poses to the corresponding target, frame-by-frame.

2.2.6 Other Advantages of Human Poses

In this section, additional advantages of using human poses as a source of information are provided. In particular, this section aims to further highlight the importance that human poses might play in other contexts, i.e. face/gaze recognition and social/group activities recognition.

Face and gaze are among the most efficient social cues for human intentions [88], many works have been done on recognising them [89–93]. In all cases these works, the authors have applied their techniques in a well-constrained scenario, where the resolution of the acquired images is large enough to see clearly the face, the expressions, even the eyes of the subjects. The latter can be useful in recognising where the target is pointing its attention and can be a predictor of its intention [88]. This is exactly the purpose of visual focus attention field [36, 89–91]. However, since the scenario is not well-constrained in video surveillance applications, e.g. the resolution of the video is not so rich in detail, many works were focused on getting an ap-

proximation of gaze direction by using head and body poses. Among those, the work in [94] is noteworthy because it is based on surveillance video in a public scenario, where the proposed algorithm infers casual events such as the meeting of people in the street. In that work, gaze direction approximation is performed to reason about the intentions of the subject. Following this topic, an interesting system aimed to recognise head and body position was the one in [95], where the position and the orientation in the 3D space of head and body are estimated by exploiting the correlation between body orientation and head orientation. Other improvements have been presented in [96, 97]. The first one is a joint probabilistic approach for pedestrian head and body orientation recognition, taking into account anatomical constraints, applied to realistic traffic settings. The second is a very recent approach for joint estimation of head and body orientations for interaction purposes, exploiting cues regarding temporal consistency and taking into account occlusions, applied to low-resolution videos of public areas. In principle, all the above-mentioned problems can be nowadays tackled by using human poses. In fact, modern detectors are able to detect face and body landmarks which can clearly inform about face and gaze directions (Fig. 2.7). Therefore, by using human poses in place of old-fashioned approaches, multiple tasks can be performed with very high efficiency. For example, in principle, it is possible to simultaneously perform HAR, face/gaze recognition, body orientation estimation and tracking by exploiting the lightness of human poses data.

Another interesting approach from the past is the one firstly proposed by Vaswani et al. [100] to recognise odd activities in video surveillance using real-world data. It was based on statistical shape theory and shape analysis [101,102]. This approach could be considered as a variant of the most general techniques regarding gait recognition [103], although it opens more general applications since it is closely connected to the concept of *activity recognition*

in video surveillance. Vaswani et al. shown in [100, 104–107] that a group of people performing a specific task separately (for example, people getting off an aeroplane and walking to reach the terminal) can be observed by a stand camera and their positions in the time-space domain can be modelled as a unique *shape*. Instead of considering each person as a different object and tracking them separately, Vaswani suggested tracing a shape using each person like a landmark. Then, the authors considered how the shape changes in the time-space domain (following the changes of mutual position between people) to estimate how much the observed configuration of landmarks is odd.

Again, the Vaswani et al. ideas can be replicated by using human poses. In principle, it is possible to integrate their solutions into a common framework based on human pose detection as the first step. Therefore, tracking,

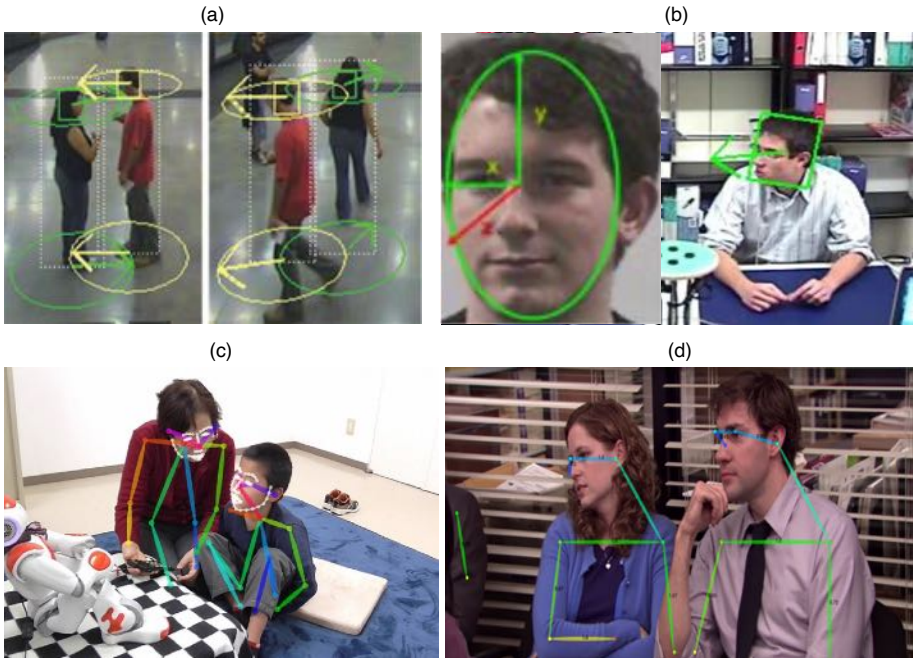


Figure 2.7: (a) Human targets and head/body orientation estimation [95]. (b) Gaze orientation recognition [89]. (c) OpenPose face and body landmarks can on two subjects, looking towards different directions [98]. (d) OpenPose body landmarks include *nose* and *hears* landmarks which can straightforwardly provide the gaze direction [99].

group activity recognition, and other above-mentioned tasks can be simultaneously performed by exploiting the efficiency of human poses.

These additional links have been not explicitly explored in this thesis. However, it has been worth mentioning them to further highlight the huge potentiality provided by human poses and to fully motivate this thesis work.

2.2.7 Human Activity Anomaly Detection

State-of-the-art HAAD methods mainly rely on RGB raw data to perform the task. In very recent work, Sultani et al. [108] presented a fully supervised RGB-based anomaly detection system. The authors defined multiple instances learning to estimate scores for each video frame, i.e. measuring the abnormality rate. Thus, in these methods, only frame-level anomalies are considered, neglecting any other semantic level, i.e. pixel-level and target-level. In semi-supervised learning, an important role is played by autoencoder based methods. Hasan et al. [109] proposed combination of CNN-based autoencoders to model the temporal evolution of both HOG and Histogram of Oriented Flows (HOF) features. Autoencoder-based approaches are based on the idea that once the autoencoder is trained on normal data only, testing data that resembles training data will be reconstructed with low error, while testing abnormalities are expected to be reconstructed with higher error. Thus, by analysing reconstruction error, it is possible to identify abnormalities. Hasan's et al. work also provide pixel-level anomaly detection. However, it still does not provide any insights of what the abnormal pixel would represent, whether a human target performing unexpected actions or a car crashing, explosions or other abnormal events. Chong et al. [110] proposed a similar approach but opting for combinations of CNN and LSTM for both spatio-temporal encoding and decoding. However, this work does not provide insights into the detected anomaly nature. Another recent approach is provided by Amraee et al. [111], where background subtraction

methods are used as a pre-processing step of raw data. This ensures that the system can focus only on moving objects. However, background subtraction methods fail on cluttered background scenes, as well as in the case of pan-tilt-zoom recording. As opposite, human pose detection is robust to these challenges. Other remarkable works [112–115] do not provide insights on how to combine HAR and HAAD and share similar common limitations such as the anomaly detection is performed at the pixel level, failing to consider which semantic region contains the pixel.

In this work, the above-mentioned autoencoder-based methods in [109] and [110] have been implemented as a state-of-the-art baseline for semi-supervised HAAD, to allow performance comparison on the proposed datasets. We demonstrate in Section 5.4.3 that these methods fail in detecting the anomalies contained in the proposed datasets.

2.3 Relevant Existing Datasets

As already mentioned throughout this thesis, HAR and HAAD challenges are closely related to the type of data is available for training, validating and test proposed solutions. In other words, different dataset proposes different challenges, which in turn encourage researches to explore HAR and HAAD from a different perspective. For this reason, in the last few decades, a considerable amount of datasets has been released. The nowadays trend is pushed by Computer Vision Community towards massive video datasets such as Kinetics [30], which includes hundreds of action classes and hundreds of thousands of video clips, showing heavily cluttered human actions. Since DL algorithms are particularly data-hungry, massive datasets are the first solution to improve performance. However, dealing with massive datasets is extremely time-consuming and power-inefficient, requiring powerful hardware platforms implementing multiple-GPU settings and parallel com-

putation, to design and optimise DL models with billions of parameters. Therefore, in recent literature, it is common to find other interesting datasets from the recent past used as benchmark [116].

Table 2.2: Relevant HAR datasets summary.

Dataset	Weizmann	KTH	i3DPost	IXMAS	UCF101	HMDB51
Actions	10	6	13	11	101	51
Samples	90	2391	832	15840	13320	6766
Targets	9	25	8	12	variable	variable
Viewpoints	1	6	8	5	undefined	undefined
Camera Motion	no	yes	no	no	yes	yes
Moving Objects	no	no	no	no	yes	yes
Cluttered Background	no	no	no	no	yes	yes
Occlusions	no	no	no	no	yes	yes
Self Occlusions	yes	yes	yes	yes	yes	yes
Channels	RGB	Mono	RGB	RGB	RGB	RGB
Resolution	180x144	160x120	960x540	390x291	342x256	352x240
FPS	25	25	25	19	variable	variable
Spontaneity	posed	posed	posed	posed	in-the-wild	in-the-wild

This thesis aims to leverage existing, famous and relevant datasets, which fits the considered problem of posture-related HAR and HAAD. In particular, this thesis focused on Weizmann [6], KTH [7], i3DPost [8], IXMAS [9], UCF101 [31] and HDMB51 [117]. In Table 2.2, a summary of the relevant dataset is provided. From a simple visual inspection of the considered datasets, it turns out that Weizmann, KTH, i3DPost and IXMAS seems to be specifically designed for posture-related HAR. In fact, the full human body is always visible and the action semantic only regards body postures, e.g. *walking*, *running*, *checking-watch*, and *bending*. Moreover, actions do not require an object to be performed. In contrast, UCF101 and HMDB51 show more complex semantic actions, which are not necessarily always performed with the same body posture. For example, the action *playing-violin*, available on UCF101, can be in principle performed sitting on a chair or standing. Table 2.2 shows that these datasets differ in terms of several aspects. In particular, as opposed to Weizmann, KTH, IXMAS and i3DPost, it is worth mentioning that: 1) UCF101 and HMDB51 present multi-target data with a single

action label for each video clip; 2) UCF101 and HMBD51 presents video clips where the camera proximity partially occludes the human bodies. In Chapter 4, these drawbacks have been detailedly explored, and quantitative comparisons between all the above-mentioned datasets have been performed, to assess each dataset suitability for posture-related HAR.

Furthermore, to properly challenge the methods proposed in this thesis,

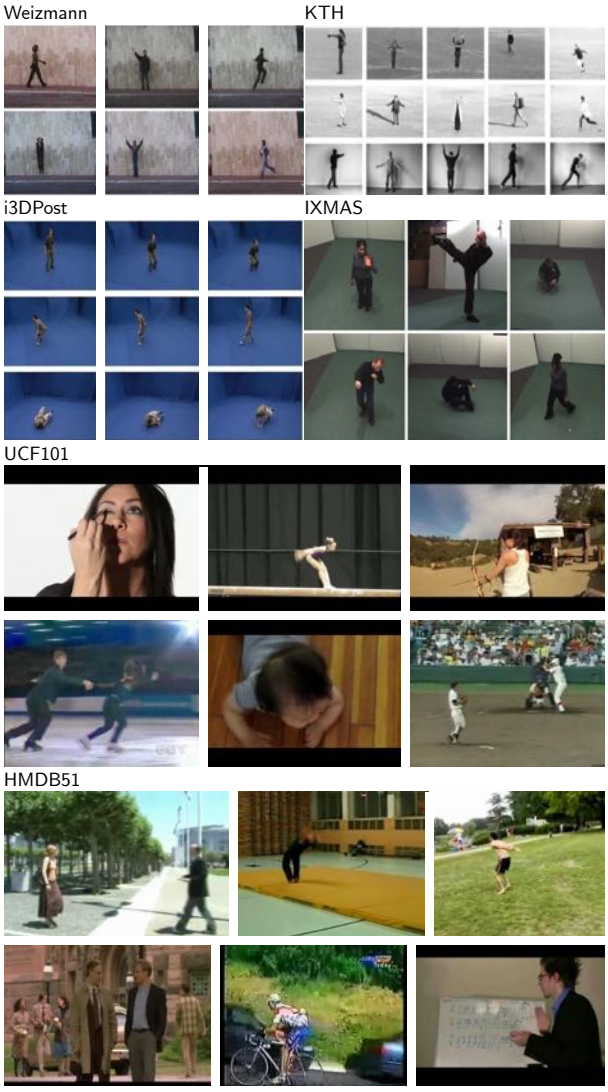


Figure 2.8: Relevant HAR dataset examples. Weizmann [6], KTH [7], i3DPost [8] and IXMAS [9] are particularly useful for posture-based HAR, since the full human body is always visible and the action classes do not involve any object. In contrast, UCF101 [31] and HMDB51 [117] are suitable for more contextually challenging HAR.

the Intelligent Sensing Lab that hosted this project has also been exploited for extensive new data recordings. In the contribution chapters, details about newly recorded data are provided.

Regarding HAAD, the literature provided several dataset which can be potentially used for HAAD, e.g. UMN [118], UCSD Ped 1&2 [114], Avenue [119], Subway Entrance and Exit [120], BOSS [121], and UCF-Crime [108]. Table 2.3 compares these datasets in terms of the number of video clips, time length, crowd activity, proposed normal and abnormal activities and events. To the best of the knowledge, these datasets are the most relevant for this project. However, as summarised in the last column of Table 2.3, from qualitative analysis, it turns out that these datasets cannot be used for this project, mainly due to semantic drawbacks. In particular, UCSD, Avenue, and Subway do not present anomalies in terms of body postures. In contrast, regarding UMN, despite it proposes normal data showing people walking and abnormal data showing people running, the video resolution is extremely low, which compromises the pose detectors functioning. Similarly, the BOSS dataset could have been potentially used, because the proposed human actions and anomalies are more similar to those considered in this thesis. However, the provided multi-target annotations are not referred to standard tracking-based ground truth. This limits BOSS applicability, unless extremely time expensive further annotation is performed to analytically assess *which* target, with *which* tracking identity, performs *which* action and *which* normal/abnormal activity. Last but not least, UCF-Crime is a popular and recent dataset which is the standard nowadays for RGB-based anomaly detection. However, most of the proposed anomalies are not posture-based, and a considerable amount of non-human abnormal events are also considered, such as explosions and car crashes. For the above-mentioned reasons, new data was required as a benchmark for the challenging HAAD considered in this project. In Chapter 5, details about the novel data

recording is provided. However, the lack of data availability demonstrates, once again, that this project addresses novel HAAD challenges, which, to the best of knowledge, are not addressed by other works in literature.

Table 2.3: HAAD Datasets Summary. State-of-the-art datasets for HAAD problem are compared in terms of number of video clips, time length, crowd activity, showing which normal and abnormal activities they propose. Moreover, in the last column, the major drawback which compromises dataset implementation is provided.

Dataset	Clips	Time	Crowd	Normal Activity Examples	Abnormal Activity Examples	Major Drawback
UMN [118]	5	5m	yes	walking	running	very limited anomalies, extremely low human body resolution
UCSD Ped 1 [114]	70	5m	yes	people walking	skater, cart, rider	no posture-related actions and anomalies
UCSD Ped 2 [114]	28	5m	yes	people walking	skater, cart, rider	no posture-related actions and anomalies
Avenue [119]	37	30m	no	people walking	throwing paper, unusual camera proximity, running	no posture-related actions and anomalies
Subway Entrance [120]	1	1h30m	no	people entering gate	walking in wrong direction, skipping payment	no posture-related actions and anomalies
Subway Exit [120]	1	1h30m	no	people exiting gate	walking in wrong direction, skipping payment	no posture-related actions and anomalies
BOSS [121]	12	27m	no	walking, sitting, standing	laying down, fighting, helping, hugging	no tracking based annotations
UCF-Crime [108]	1900	128h	no	traffic, walking, standing	abuse, arrest, assault, accident, burglary, explosion, fighting, robbery, shooting, shoplifting	no posture-based actions and anomalies, multiple non-human events anomalies

2.4 Performance Measures

In this section, standard methods to measure classification and anomaly detection performance are summarised. These metrics are used to assess the performance of the proposed algorithms, and to compare the obtained results with the state-of-the-art.

Let consider a binary classifier, which can only return a Positive (**P**) and a Negative (**N**) output. Therefore, as shown in Table 2.4, the most common quantities which can be computed are the following:

- True Positive (**TP**): number of **P** testing samples which are correctly classified;
- False Negative (**FN**): number of **P** testing samples which are incorrectly classified;
- True Negative (**TN**): number of **N** testing samples which are correctly classified;
- False Positive (**FP**): number of **N** testing samples which are incorrectly classified;

The standard metric for a binary classifier is the *accuracy* A, which is defined as

$$\mathrm{A} = \frac{\mathbf{TP} + \mathbf{TN}}{\mathbf{TP} + \mathbf{FN} + \mathbf{TN} + \mathbf{FP}} = \frac{\mathbf{TP} + \mathbf{TN}}{\text{Total Testing Samples}} \tag{2.4.1}$$

Table 2.4: Confusion matrix for a binary classifier.

		Predicted **P**	Predicted **N**
Ground Truth	**P**	**TP**	**FN**
	N	**FP**	**TN**

Confusion Matrix

In the case of multi-class classification, let suppose that a classifier is trained on classes $\mathcal{L} = \{l_1, \ldots, l_m\}$. The common generalisation for the above-mentioned binary metric is provided by the multi-class *confusion matrix* C. An example of C is provided in Table 2.5-(a). By definition, entries $c_{i,j} \in$ C correspond to the number of testing samples of class l_i have been classified as l_j. Therefore, as in Table 2.4, the rows of C correspond to the ground truth classes and the columns correspond to the labels predicted by the classifier. In particular, the confusion matrix C in Table 2.5-(a) is *absolute*, in the sense that reported numbers are not scaled to the total number of testing samples for each class. Conversely, in Table 2.5-(b), the *relative* confusion matrix $\bar{\text{C}}$ is computed by dividing each entry of the absolute confusion matrix by the total number of testing sample in the corresponding class.

Therefore, the $\bar{\text{C}}$ can give normalised information in the case of *unbalanced* class, i.e. the number of total samples for each class is uneven.

Given the multi-class confusion matrix, the accuracy in (2.4.1) can be generalised as follows:

$$\text{A} = \underbrace{\frac{\sum_{i=1}^{m} c_{i,i}}{\sum_{i=1,j=1}^{m,m} c_{i,j}}}_{\text{defined by using C}} = \underbrace{\frac{1}{m}\sum_{i=1}^{m} \bar{c}_{i,i}}_{\text{defined by using } \bar{\text{C}}} \tag{2.4.2}$$

ROC Plot

The ROC Plot is graphical method to illustrate binary classifiers performance. Typically, in case of binary classification, the predicted class l_1 or l_2 is provided in terms of probabilities of score functions, as already discussed, for example, in (2.1.7) for the logistic regression. In particular, given a sample $\boldsymbol{x}_i$, its predicted class label $L(\boldsymbol{x}_i)$ normally depends on a threshold T, as

follows:

$$L(\boldsymbol{x}_i) = \begin{cases} l_1 & \text{if} \quad \mathbf{P}(L = l_i | \boldsymbol{x}_i, \boldsymbol{w}) > T, \\ l_2 & \text{otherwise} \end{cases} \tag{2.4.3}$$

Therefore, by varying the threshold T, the performance of the classifier changes. Let suppose that TPR(T) and FPR(T) are, respectively, the true positive and the false positive rates for the given binary classifier, with respect to the threshold T. By definition, TPR $= \frac{\text{TP}}{\text{TP+FN}}$ and FPR $= \frac{\text{FP}}{\text{FP+TN}}$. Therefore, the ROC curve is obtained by plotting TPR(T) against FPR(T). An example of ROC curve is provided in Fig. 2.9. ROC curves are effective tools to compare multiple classifiers. In general, the more the ROC curve is close to the upper left corner, i.e. TPR = 1 and FPT = 0, the better.

Table 2.5: Confusion matrix for a multi-class classifier. (a) Absolute confusion matrix C, reporting the absolute number of samples predicted for each class, where the total number of testing samples for each class is set to be, for example, 20. (b) Relative confusion matrix $\bar{\text{C}}$, reporting the prediction rate for each class, obtained by dividing each entry of C by 10.

(a) C

		Predicted				
		l_1	l_2	...	l_{m-1}	l_m
Ground Truth	l_1	20	0	...	0	0
	l_2	0	20	...	0	0
	...	...	...	...	...	...
	l_{m-1}	0	0	...	17	3
	l_m	0	2	...	5	13

(b) $\bar{\text{C}}$

		Predicted				
		l_1	l_2	...	l_{m-1}	l_m
Ground Truth	l_1	1	0	...	0	0
	l_2	0	1	...	0	0
	...	...	...	...	...	...
	l_{m-1}	0	0	...	.85	.15
	l_m	0	.1	...	.25	.65

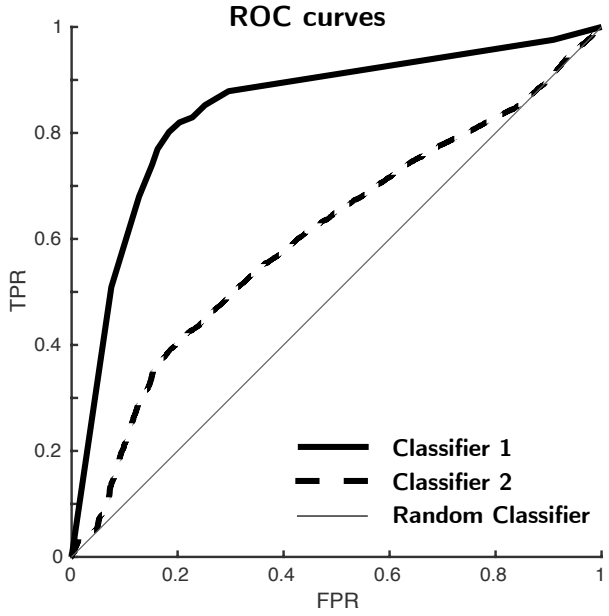


Figure 2.9: ROC curves for two classifiers, i.e. Classifier 1 and Classifier 2. The Classifier 1 outperforms the Classifier 2. The Random classifier is the worse possible binary classifier, which predict classes following a Bernoulli distribution with parameter $p = 0.5$.

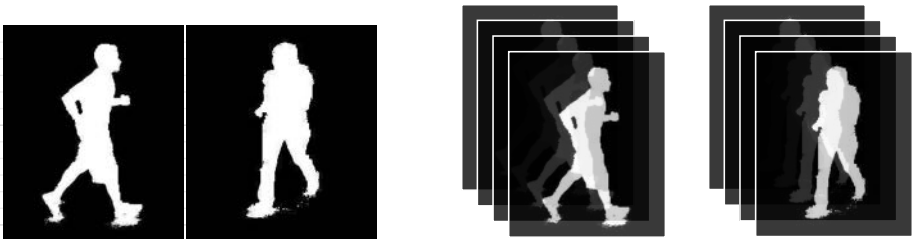
Chapter 3

3D-HOG EMBEDDING FRAMEWORKS FOR SILHOUETTES-BASED HAR

3.1 Introduction

Human silhouette plays a key role in posture-based HAR since it carries the major part of the posture-related information. Human silhouettes are normally provided in the form of binary masks, as shown in Figure 3.1-(a), where the background is represented by 0, while the foreground target (human) is represented by 1. Therefore, only information about the human body shape is preserved. However, considering that multiple action frames might be available, it is possible to stack multiple masks and obtain a spatio-temporal action representation based on human silhouettes, i.e. Space-Time Volume (STV), as defined in Section 2.2.2. Therefore, the STV can be formally represented as a 3D binary matrix, where the first two dimensions represents the *width* and *height* of the target ROI, while the third dimension represents the *time*, as shown in Figure 3.1-(b).

As already discussed in Section 2.2.2, STVs are common action representations for silhouette based HAR, among other representations, such as motion-history maps. Moreover, when STVs are used, 3D-HOG features can



(a) (b)

Figure 3.1: (a) Examples of human silhouettes obtained via background subtraction. (b) Space-Time Volumes examples obtained by piling up several action frames.

be extracted from each local region of the STV. In this chapter, the Weinland et al. framework [26] for 3D-HOG based HAR has been implemented as a baseline. Since no public code was available, the baseline has been replicated following the original paper instructions [26, 43, 44]. Regarding the 3D-HOG features extraction implementation, it has been reproduced by following instructions in [47,122]. Therefore, results published in [26] regarding Weizmann dataset have been successfully replicated.

The above-mentioned work allowed to highlight limitations of the Weinland et al. framework, which motivated this chapter contribution. In particular, this chapter contribution is twofold: 1) the proposed frameworks are more accurate and stable over different training rounds than the baseline; 2) the proposed frameworks also outperforms other state-of-art methods in terms of recognition accuracies and robustness to appearance changes over the tested datasets. These results have been obtained by addressing Objectives 1 and 2 mentioned in Section 1.3.

The remaining of this chapter is organised as follows. In Section 3.2, the problem, the 3D-HOG feature extraction, the baseline embedding algorithm and its limitations, are discussed. In Section 3.3, the proposed frameworks are presented. In Section 3.4, simulations and results obtained on Weizmann and i3DPost datasets are reported, showing that the proposed frameworks outperform the baseline and other approaches. In Section 3.5, Objective 3

in Section 1.3 is also addressed, providing a critical evaluation on the 3D-HOG frameworks effectiveness in real-world applications and on potentially disruptive issues that might affect practical implementation. Finally, in Section 3.6, the conclusions of this chapter are drawn.

3.2 Preliminaries

3.2.1 3D-HOG Feature Extraction

Let us consider s as defined in Section 1.3.1 and be in the form of a STVs. a STV partition made by defining *overlapping blocks* of a fixed dimension. Binary data within each block is then used to compute the 3D-HOG descriptor as suggested in [26], exploiting the 3D vectorial gradients field. The 3D-HOG feature extraction is summarised in Figure 3.2. Each STV block is further partitioned into non-overlapping cubic cells C. Therefore, the mean gradient is computed over the binary field contained within each cell. In particular, the mean gradient $\bar{g}$ is computed as $\bar{g} = \frac{1}{|C|}\sum_{x,y,t} g_{x,y,t}$, where $|C|$ is the cardinality of the each cubic cell and $g_{x,y,t}$ is the punctual gradient computed in the position (x, y, t). Therefore, the magnitude of $\bar{g}$ is quantised according to a pre-defined Platonic solid. In particular, $\bar{g}$ is projected onto the solid faces unit vectors and the projection magnitude is computed. Thus, the obtained magnitudes are considered in a pre-defined and fixed order and represented as an histogram. Finally, cell histograms are concatenated, i.e. histograms are listed as a series following a SIFT-like [122] approach, to obtain a robust block descriptor. Following the above-mentioned procedure, let $\mathbf{b}$ be the 3D-HOG block descriptor. This descriptor can be considered highly informative with respect to the 3D-shape defined by binary data within the block [26], as well as robust to noise and small data deformations.

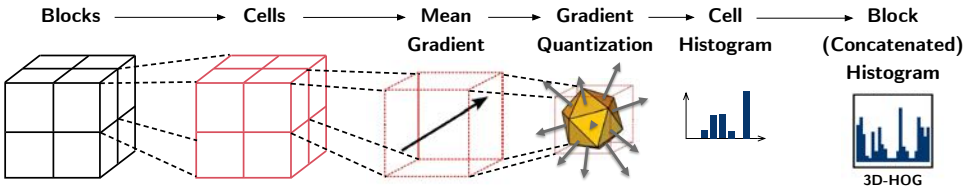


Figure 3.2: 3D-HOG feature extraction pipeline. STV overlapping blocks are further partitioned into non-overlapping cells. Therefore, the mean gradient is computed within each cell and quantised according to a pre-defined solid. Obtained quantisation is represented as histogram. Thus, cell histograms are concatenated to obtain the block 3D-HOG descriptor.

3.2.2 Prototypes Library Embedding

The block descriptors $\mathbf{b}$ obtained in the previous section are organised as follows. In this chapter, let $p = (r_0, c_0)$ be a space position in the STV. Thus, the *flow at location* p is defined as $\mathcal{B}_p = \{\mathbf{b}_{1,p}, \mathbf{b}_{2,p}, \ldots, \mathbf{b}_{K,p}\}$, where $\mathbf{b}_{i,p}$ is the ith block descriptor at location p and K is the total number of block descriptors in the time dimension. Different locations p can be considered, depending on the chosen space partition. Without loss of generality, let $p \in \{p_1, \ldots, p_P\}$, where P is the number of considered location in the (r, c) plane. In Figure 3.3, an overview of the STV and the blocks partitioning is shown.

Following the strategy in [26], for each fixed label l and point of view w, n block descriptors are randomly chosen with respect to time and space within

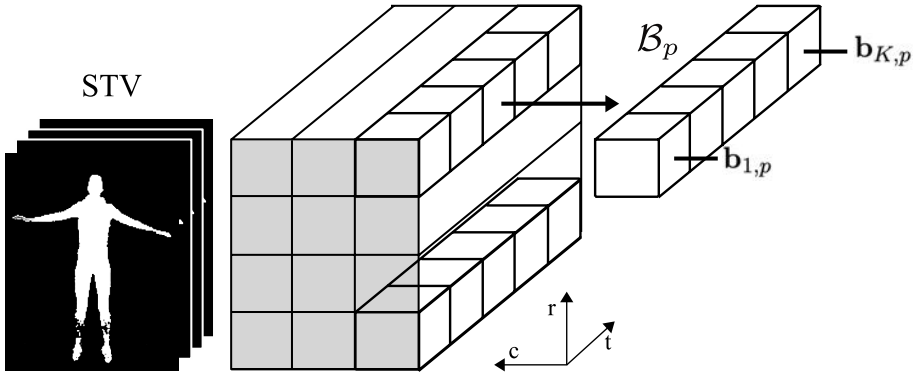


Figure 3.3: STV and its overlapped blocks partitioning. The flow $\mathcal{B}_p$ is depicted, for a generic position $p = (r_0, c_0)$.

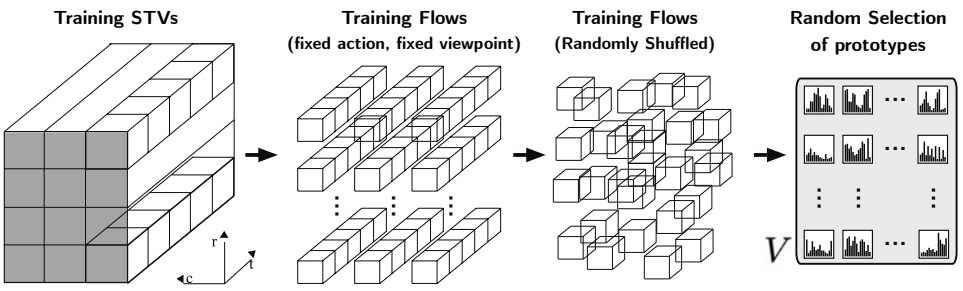


Figure 3.4: Random library selection process, as suggested by Weinland et al. in [26].

the training STVs. This yields to a descriptors library V such that $|V| = nLW$. Figure 3.4 visually shows this strategy. Without loss of generality, $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_{nLW}\}$, where it is specified that V contains n descriptor prototypes for each action and each point of view included in the dataset. It is crucial to mention that, in [26], the random selection of prototypes is performed *across* the whole STV, without considering the position p. Therefore, each prototype in V belongs randomly to any flow $\mathcal{B}_p$ provided by training data.

As illustrated in Figure 3.5, the scalar *embedding* $D_i(\mathcal{B}_p)$ of each block flow $\mathcal{B}_p$ against the library V is defined as follows:

$$D_i(\mathcal{B}_p) = \min_{j=1,\ldots,K} d(\mathbf{b}_{j,p}, \mathbf{v}_i) \qquad i = 1, \ldots, nLW \tag{3.2.1}$$

where d represents the Euclidean distance. Thus, the *embedded vector* $\mathcal{D}_p$ at location p is defined as

$$\mathcal{D}_p = [D_1(\mathcal{B}_p), \ldots, D_{nLW}(\mathcal{B}_p)] \tag{3.2.2}$$

It is worth underlining that, since the library V explicitly contains some block descriptors selected from training data, the embedded vector in equation (3.2.2) might contain some zero-entries. This occurs when the flow $\mathcal{B}_p$, which is going to be embedded, actually contains at least one of the

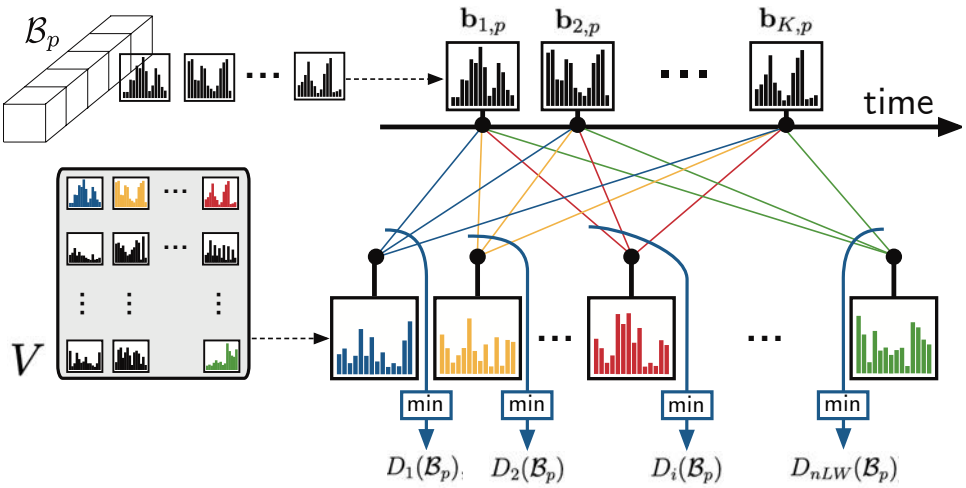


Figure 3.5: 3D-HOG prototypes embedding, as suggested by Weinland et al. in [26].

prototypes in V.

3.2.3 Final Flow-based Decision

For each location p, a classification model Θ_p is learned by using embedded vectors $\mathcal{D}_p$ obtained with training data. In particular, following the approach in [26], L_2-Regularised Logistic Regression model is trained. Thus, the probability $\mathbb{P}(l_p^* = l|\mathcal{D}_p^*, \Theta_p)$, of a testing embedded vector $\mathcal{D}_p^*$ to belong to one of the considered action classes l is obtained.

Depending on the chosen space partition, this strategy leads to *flow-based* independent decisions in the number of P. To combine all these decisions to classify the action in the STV, the *Sum Rule* has been considered, as in [26], since it is easy to implement and effective. Moreover, the Sum Rule is claimed to be more rewarding in terms of accuracy [26]. Formally, the *Sum Rule* consists of selecting the final label l^* such that

$$l^* = \max_{l \in \mathcal{L}} \sum_{p=1}^{P} \mathbb{P}(l_p^* = l|\mathcal{D}_p^*, \Theta_p) \tag{3.2.3}$$

In Figure 3.6, the pipeline of the baseline is provided.

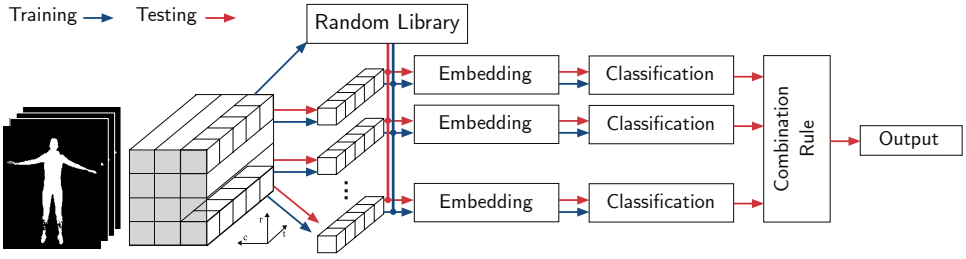


Figure 3.6: Baseline pipeline. A cross-location library is created by randomly sampling the training 3D-HOG prototypes. Thus, for each location, the correspondent block flow is embedded by using the library. Training embedded vectors are used to train a local classifier. Therefore, testing local embedded vectors are classified accordingly. A non-trainable combination rule is used to combine the local confidences in order to obtain the final action label.

3.2.4 Baseline Limitations

The implementation of the baseline highlighted important limitations which motivated the contributions of this chapters. In particular, two major limitations are discussed in this section.

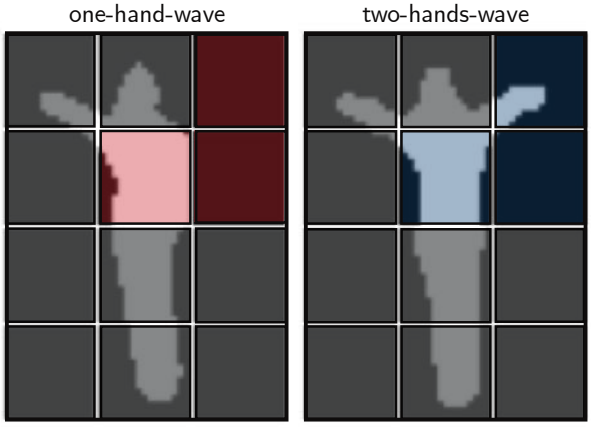


Figure 3.7: The baseline performs independent, location-based classifications, which are later combined with a simple combination rule based on posterior probabilities. Therefore, in case of similar actions such as *one-hand-wave* and *two-hands-wave*, the local classifiers might make mistakes in those locations where the two actions look very similar (grey squares). As opposed, only in discriminative locations (red/blue squares) there are high chances to obtain good outcomes.

Action Label-based, Independent, and Local Classifiers

The baseline performs several independent, location-based classifications, depending on the number of chosen locations in the STV. In the testing phase, the output of these classifiers (posterior probabilities) are combined with a simple combination rule, e.g. Sum Rule, *which does not depend on trainable parameters*. To highlight how this represents a limitation, let us consider two similar actions such as *one-hand-waving* and *two-hands-waving* as in Figure 3.7. Let us also consider classifiers trained on positions where the two actions appear similar (grey squares in Figure 3.7). It is reasonable to argue that these classifiers will output unreliable responses. Therefore, since several positions are affected by this problem, there might be a high number of unreliable contributions which can compromise performance. The only classifier responses which are reasonably reliable are those obtained in highly discriminative locations (red/blue squares in Figure 3.7). However, since there might be few discriminative positions, their impact on the final output is limited when non-trainable combination rules are adopted. In other words, independently trained classifiers and the untrainable decision rule which combines them all, make the baseline suboptimal.

Performance Instability

The definition of V is based on a random selection of prototypes. This choice introduces consistent variability in the training phase. In fact, different random choices might change the trained model. Thus, different training round might end up with different models, which, in turn, might show different testing performance. This instability can be potentially mitigated by increasing the number of prototypes in V, i.e. increasing n. However, in Section 3.4, it is shown that this is not necessarily the case, and other strategies are needed to avoid the random selection prototypes.

3.3 Proposed Frameworks

The goal of this section is to describe the proposed modifications to the baseline, to overcome the above-mentioned limitations. In Section 3.4, it is demonstrated that not only these limitations are resolved, but also that the proposed methods yield to more accurate and robust performance.

3.3.1 Robust Prototypes Selection

The first modification regards the definition of the prototype library V. In this section, the cross-location random selection of prototypes is replaced with a location-based selection strategy. Let $H = H(p, l, w)$ be the total number of descriptors with the same labels l and w in $\mathbb{T}$ at location p. Therefore, the set of training block descriptors corresponding to the given p, w and l are $\{\mathbf{b}_{j,p}\}_{j=1}^{H}$, where subscripts l and w have been omitted for simplicity. Since each block descriptor can be seen as a point within a multidimensional cartesian space, the Hierarchical Clustering algorithm can be deployed to seek the internal cluster structure of training blocks $\{\mathbf{b}_{j,p}\}_{j=1}^{H}$. Therefore, n clusters of descriptors can be computed such that:

$$\{\mathbf{b}_{j,p}\}_{j=1}^{H} = \{\mathbf{b}_{j,p}\}_{j=1}^{S_1} \cup \{\mathbf{b}_{j,p}\}_{j=S_1}^{S_2} \cup \cdots \cup \{\mathbf{b}_{j,p}\}_{j=S_{n-1}}^{S_n} \tag{3.3.1}$$

where S_1 is the cardinality of the first cluster, $S_2 - S_1$ the cardinality of the second cluster and similarly $S_n - S_{n-1}$ is the cardinality of the last cluster. In (3.3.1), action and point of view symbols are implicit, since all descriptors have fixed l and w. Thus, by computing the cluster centre, i.e. by averaging elements within the same cluster, n *new* prototypes remain defined.

By varying all labels l and w, this definition leads to a new library $\bar{V}$ such that $|\bar{V}| = nLW$ prototypes. Due to the proposed definition, prototypes in $\bar{V}$ are no longer included within the training subsequences. As a consequence, the embedded vectors, defined by using (3.2.1) and (3.2.2), will not have zero-

entries, which helps in preventing overfitting. In Section 3.5, further details about this topic are provided. Moreover, as experiments presented in Section 3.4 show, this deterministic strategy stabilises the training process, since the random selection is replaced by a deterministic process. Moreover, as shown in Figure 3.8, it is reasonable to argue that clusters-based prototypes tend to be more informative and exhaustive than prototypes chosen by random selection.

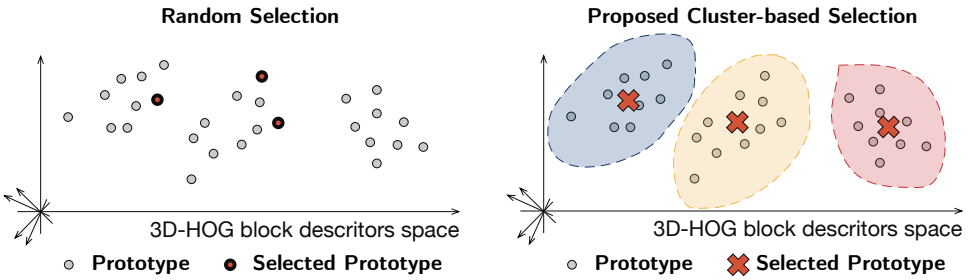


Figure 3.8: Comparison between the baseline, random-based, 3D-HOG prototype selection and the proposed, clusters-based, 3D-HOG prototype selection for V. Due to the random selection, internal cluster distributions are not be taken into account by the baseline method. Moreover, different random selection might lead to considerably different prototype libraries. In contrast, the proposed method considers internal clusters to compute new prototypes, providing a deterministic strategy that does not change over training rounds.

3.3.2 Overcoming Flow-based Decisions Rules

In this section, two novel approaches for combining local embedded vectors for a cross-position classification are proposed.

Proposed Framework A

Given a fixed location p, the training embedded vectors can be seen as points within a multidimensional space. Thus, by exploiting only training data labels l, the following *training* process is proposed:

1. Exploiting training embedded vectors $\{\mathcal{D}_{p,i}\}_{i=1}^{|\mathbb{T}|}$, L_2-Regularised Logistic Regression can learn a positional model Θ_p. Therefore, Θ_p maps testing data to a temporary label $\tilde{l}_{\mathcal{D}_p}$;

2. Exploiting training embedded vectors $\{\mathcal{D}_{p,i}\}_{i=1}^{|\mathbb{T}|}$, PCA can simultaneously reduce the dimensionality of $\{\mathcal{D}_{p,i}\}_{i=1}^{|\mathbb{T}|}$. Thus, by averaging the PCA-transformed vectors according to their class label, the center class points $\bar{\mathbf{C}}_p = \{\bar{\mathbf{c}}_{p,l}\}_{l\in\mathcal{L}}$ can be computed. Therefore, $\bar{\mathbf{C}}_p$ is made by vectors with a small and fixed number of components which corresponds to an explained variance α_1.

3. Let $\mathcal{D}_p$ be a specific training vector and $\tilde{l}_{\mathcal{D}_p}$ be the temporary label provided by model Θ_p by performing the step 1. Therefore, $\mathcal{D}_p$ can be associated with $\bar{\mathbf{c}}_{p,\tilde{l}_{\mathcal{D}_p}}$. Thus, by varying the position p, each training sample s_i can be associated with a single vector by concatenating vectors $\bar{\mathbf{c}}_{p,\tilde{l}_{\mathcal{D}_p}}$ for all p, that is $\bar{\mathcal{C}} = \{[\bar{\mathbf{c}}_{1,\tilde{l}_{\mathcal{D}_1}}, \ldots, \bar{\mathbf{c}}_{P,\tilde{l}_{\mathcal{D}_P}}]\}_{l\in\mathcal{L}}$. Therefore, a new L_2-Regularised Logistic Regression model can be trained, to get a cross-locations model Θ.

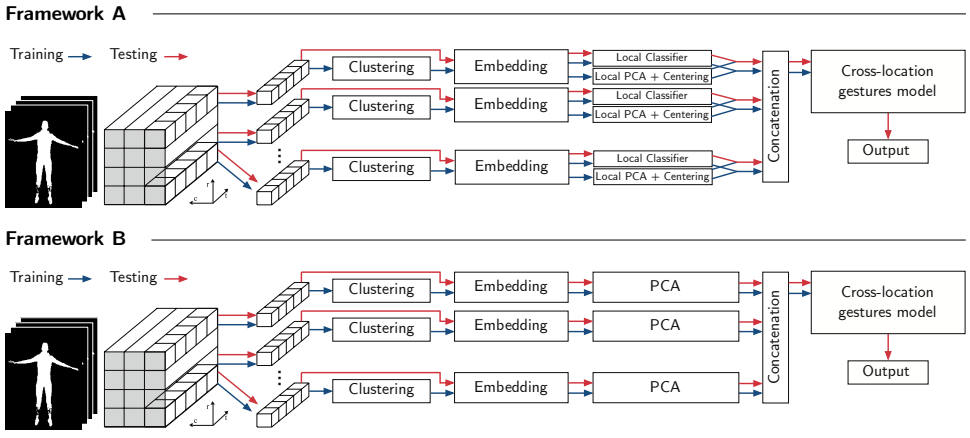


Figure 3.9: Framework A and B pipelines. Both frameworks share the same pre-processing, i.e. clustering followed by embedding. Similarly, both frameworks concatenate local features to train a cross-location gestures model. However, they differ on the way the local features are computed.

A testing embedded vector $\mathcal{D}_p^*$ is labelled with a positional label $\tilde{l}_{\mathcal{D}^*_p}$ by using the learnt model Θ_p and associated with $\bar{\mathbf{c}}_{p,\tilde{l}_{\mathcal{D}^*_p}}$. Thus, the testing sample will be associated with a single vector given by concatenating $\bar{\mathbf{c}}_{p,\tilde{l}_{\mathcal{D}^*_p}}$ for all p, that is $\bar{\mathcal{D}}^* = [\bar{\mathbf{c}}_{1,\tilde{l}_{\mathcal{D}^*_1}}, \ldots, \bar{\mathbf{c}}_{P,\tilde{l}_{\mathcal{D}^*_P}}]$.

The final label l^* is obtained by using Θ to map $\bar{\mathcal{D}}^*$ to l^*. In Figure 3.9, the training and testing pipeline for this case is shown. It is worth specifying that the parameter α_1 in the training process can be experimentally be established. Normally, α_1 can be set to 95% or 99%. However, the higher the parameter, the heavier the training of the model Θ, since the higher the number of considered features. However, since not necessarily more features correspond to better results as overfitting and curse of dimensionality might occur, α_1 is requested to be the smallest that ensures good performance.

Proposed Framework B

In this section, the above-mentioned framework pipeline is simplified.

For a given location p, the training process defined in the previous section can be reduced to a PCA over the training embedded vectors $\{\mathcal{D}_{p,i}\}_{i=1}^{|\mathbb{T}|}$, considering the number of principal components that reaches the cumulative α_2 percentage of explained variance. Let $\{\bar{\mathcal{D}}_{p,i}\}_{i=1}^{|\mathbb{T}|}$ be the transformed training embedded vectors and $\mathbf{A}_p$ the transformation matrix. Therefore, each training sample s_i will be associated with a single vector given by concatenating $\bar{\mathcal{D}}_{p,i}$, for all p.

In the testing phase, the testing embedded vector $\mathcal{D}_p^*$ can be centred and transformed by using $\mathbf{A}_p$, according to

$$\mathbf{A}_p \left(\mathcal{D}_p^* - \frac{1}{|\mathbb{T}|} \sum_{i=1}^{|\mathbb{T}|} \mathcal{D}_{p,i} \right) = \bar{\mathcal{D}}_p^* \tag{3.3.2}$$

Therefore, testing samples are associated with a single vector given by concatenating $\bar{\mathcal{D}}_p^*$ for all p. Training and testing data can finally supply an L_2-Regularised Logistic Regression, which in turn provide the final label l^* for testing data. In Figure 3.9, the training and testing pipeline for this case is shown.

Similarly to α_1, the parameter α_2 is requested to be sufficiently small

to avoid the curse of dimensionality, but sufficiently high to allow adequate data description.

3.4 Experiments

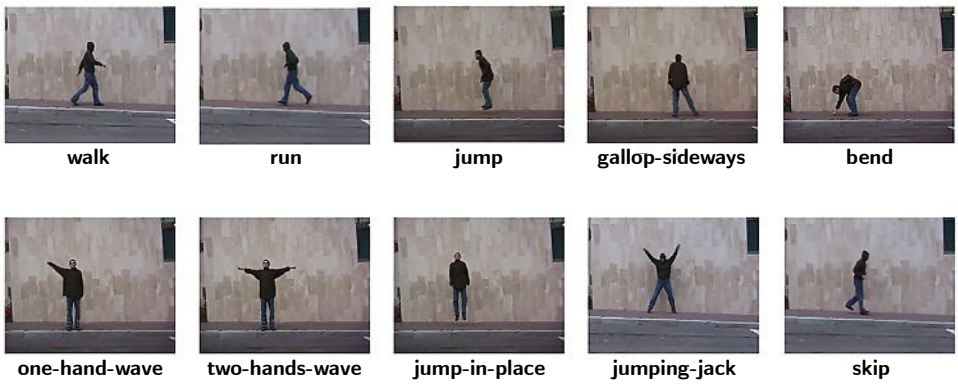


Figure 3.10: The Weizmann dataset action classes [6].

The single-viewpoint Weizmann dataset consists of video samples of 10 actions performed by 9 actors recorded from the single action-related point of views. Figure 3.10 shows Weizmann examples for the considered actions. Binary masks are publicly available in the requested STV form. Regarding the multi-viewpoints i3DPost dataset, it consists of video clips recording 6 single-actor actions, 2 multi-actors actions videos, 4 multi-action single-actor videos and facial-expressions data. The actions are performed by 8 actors and simultaneously recorded from 8 different points of view. The whole dataset was considered, ignoring video clips containing multiple-actions and facial-expressions. Figure 3.11 shows i3DPost examples for the considered actions and viewpoints. ViBE algorithm [123] was used for background subtraction, to get the binary masks of the scene. Then, the ROIs around the subjects were hand-picked exploiting the human shape centroid. As in [26], for both datasets, the STVs were rescaled from the original video size to $64 \times 48 \times t$ pixels. The block dimension was fixed to $16 \times 16 \times 16$ pixels, with overlapping of 8 pixels. For the 3D-HOG feature extraction, the

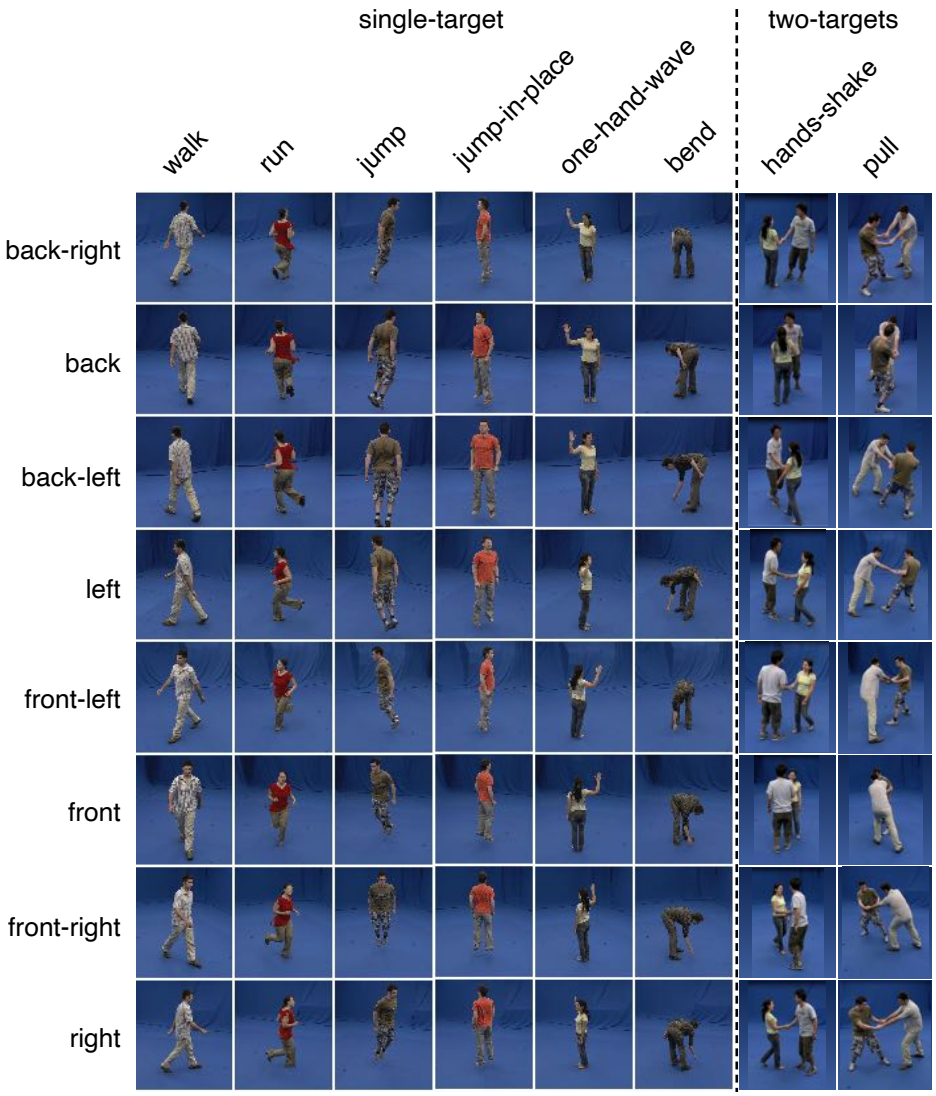


Figure 3.11: The i3DPost dataset action classes and viewpoints [8].

same setting as in [26] was used.

In Section 3.4.1, the baseline limitations are shown by explicitly depicting some samples output as an example. Therefore, in Sections 3.4.2 and 3.4.3, the main results for the proposed methods are presented, making comparison with the baseline and the state-of-the-art. In these sections, the robustness to human appearance changes was tested with the *leave-one-actor-out* (LOAO) experimental setting. Thus, one actor samples were kept out from the training and used for testing. Accuracy results are given on av-

erage over all possible LOAO configurations. The averaged accuracy results for my implementation of the baseline are provided with standard deviation σ, over ten different choices of the random prototypes in V. Regarding the proposed method, since it is based on a deterministic strategy for choosing prototypes in V, the results are fixed for each n, without any variation. In Section 3.4.4, the computational complexity of the proposed frameworks is discussed. In Section 3.4.5, interesting examples are shown and discussed to highlight how different time windows can affect performance.

3.4.1 Baseline Limitations Evidences

In this section, four samples from the i3DPost dataset are studied to highlight and confirm the expected baseline limitation discussed in Section 3.2.4. In particular, it is shown that the baseline outputs might be potentially impacted by no-weighted local classifiers outputs.

In Figure 3.12, two very similar samples are considered, i.e. *jump* and *jump-in-place*. The only visual difference between these two samples is that the jumping action is performed moving forward, while the jumping-in-place is performed remaining in the same position. Therefore, from a mechanical point of view, the two actions are almost identical, although the vertical axes of the two targets are slightly different.

The first set of mistakes it is possible to identify are related to zero-data. An *empty* position (zero-data) is expected not to provide any information. As opposed, the local classifier trained by the baseline provides an output. This output can be *unreliable* (if that position is never or always expected to be empty), *biased* (if zero-data is expected for a certain subset of action). Therefore, it would be preferable to assign a dynamic weight to zero-data positions. The dynamic weight depends on which information the other positions are carrying. However, the baseline assigns the same importance to every location, including locations that have zero data.

The second set of mistakes regards positions where the baseline simply misinterpret the performed action. For example, in both sample in Figure 3.12, in the upper-left corner of the STVs, the baseline consistently assign to *hand-wave* high scores. These mistakes lead to a wrong final estimated label in the *jump-in-place* sample. However, the local classifiers correctly recognise the actions in the region of targets' legs and backs. Therefore, for example, it would be preferable to assign a low weight to the upper-left corner decisions when the *jump* and *jump-in-place* actions are detected on the legs and back related positions.

In Figure 3.13, additional examples are depicted, to further show that the local classifiers are not performing well due to the above-mentioned mistake types. In these examples, the mistakes do not compromise the final output. However, the correctness of the estimated label is only guaranteed by the fact that the *number* and the *confidence* of correct local decisions is higher than the number and the confidence of the mistaken local decision.

I conclude that the Sum Decision Rule, which the baseline exploits to combine the local classifiers decisions, is unreliable and might lead to unpredictable outcomes. It is evident from these pictures that different positions may carry different information, which in turn is required to be weighted in the process of making the overall decision. This requirement motivates the proposed methods, which introduces a classifier that combines information coming from different locations and assigns a weight to each of them by using Logistic Regression. In this way, the assigned weights clearly dynamically depends to each other.

3.4.2 Weizmann and i3DPost (Single-Viewpoint) Results

In this section, single-viewpoint experiments are performed to compare the proposed frameworks results with the baseline and other state-of-the-art approaches in the simplest scenario.

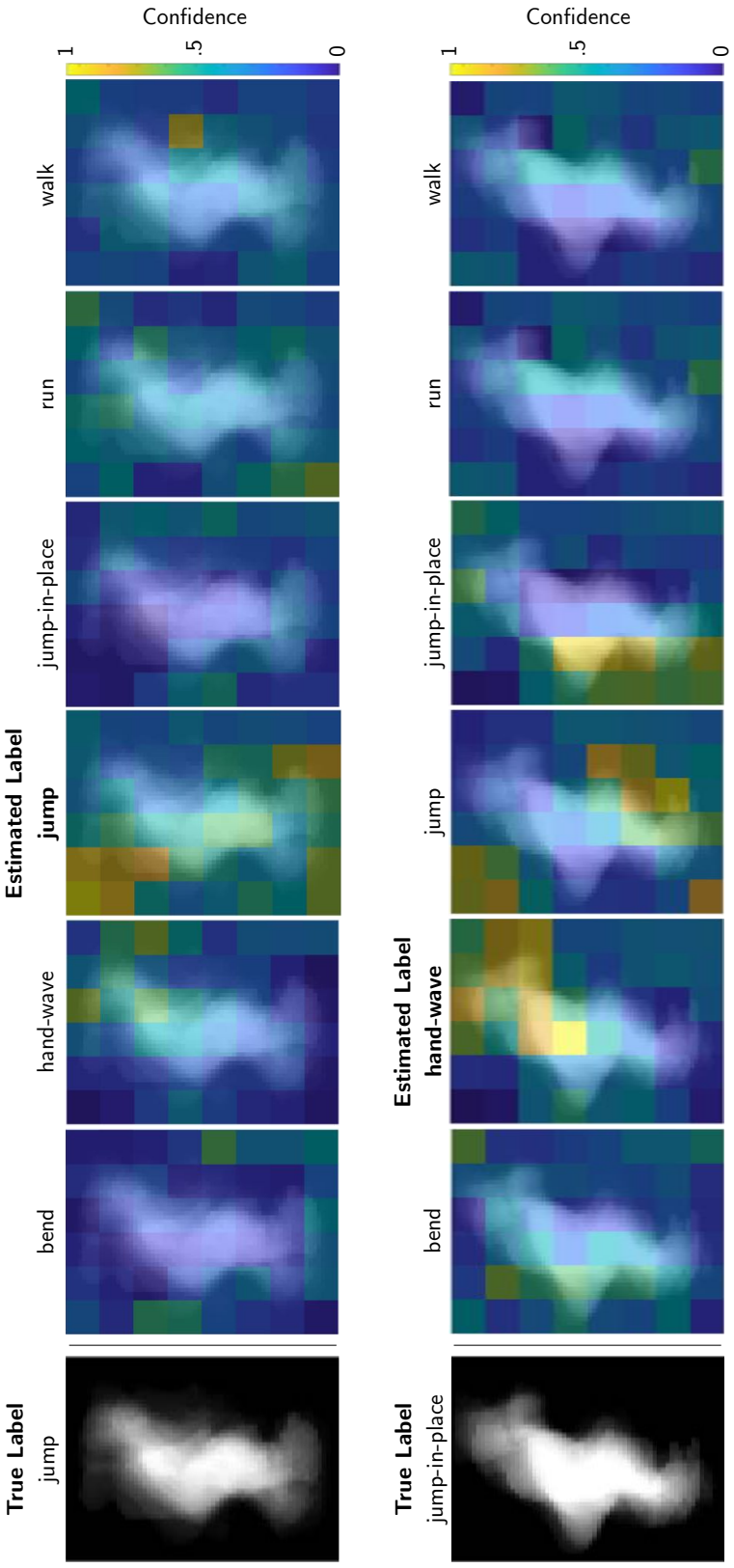


Figure 3.12: Comparison of local classifiers responses, in the case of two similar samples, i.e. *jump* and *jump-in-place*. The confidences in the same positions sum up to one. In the *jump* sample, the estimated label is correct, while in the *jump-in-place* case, the baseline outputs *hand-wave*.

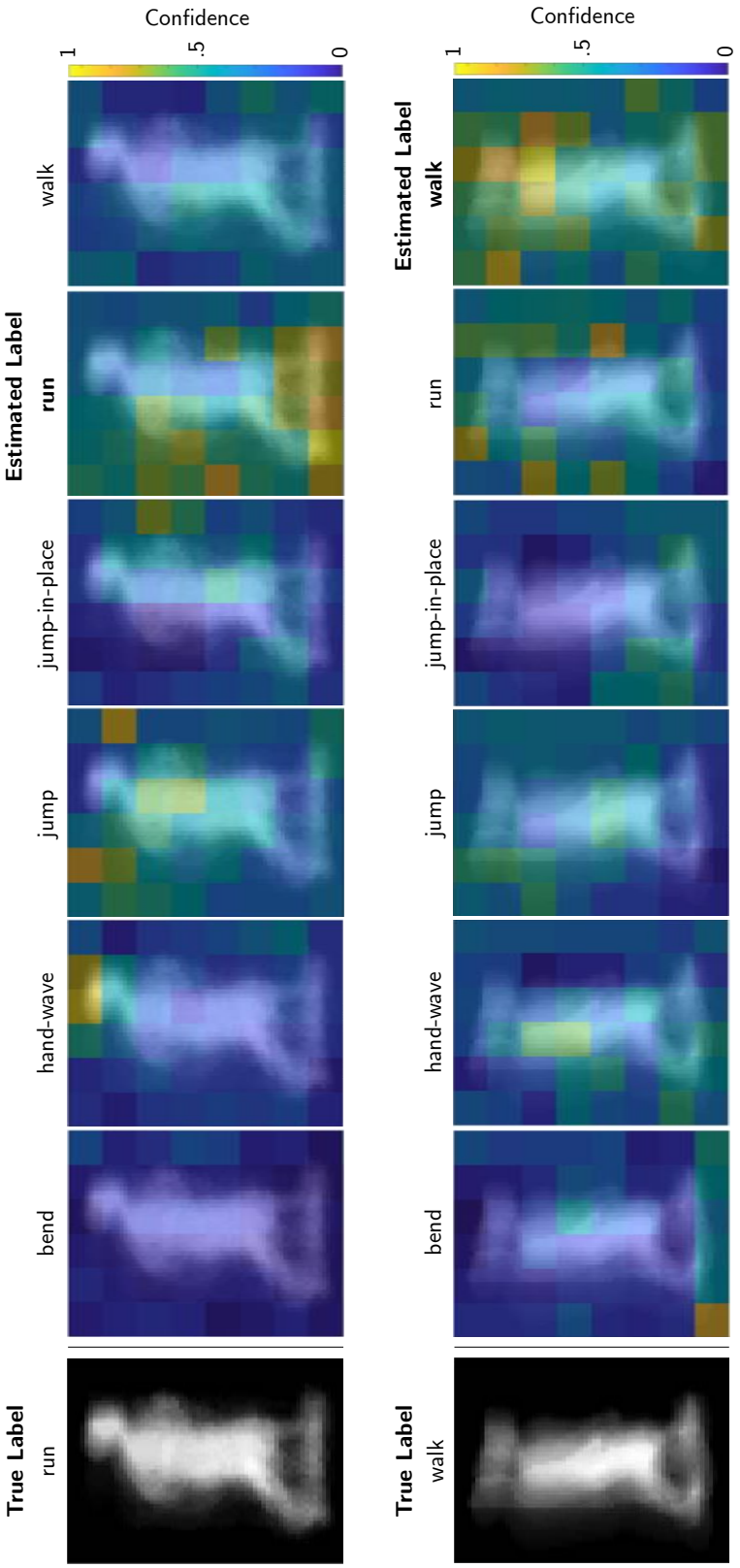


Figure 3.13: Comparison of local classifiers responses, in the case of two similar samples, i.e. *run* and *walk*. The confidences in the same positions sum up to one. In both cases, the estimated label is correct. However, in numerous positions, the classifiers make mistakes.

Regarding Weizmann dataset, my implementation of the baseline method has confirmed the perfect result (accuracy 100%) reported in [26] for 9 actions over 10 (*two-hands-wave* action out), by using $n = 30$ prototypes per action in the baseline random library. This result suggests that my implementation of [26] is correct. However, [26] does not provide results for the whole set of actions of the Weizmann dataset. My implementation of [26] shows an accuracy of at most 98.66% in this case, achieved with $n = 30$ and with a standard deviation $\sigma = 0.61$ (over 10 different training rounds). In contrast, in this case, the proposed Framework A is able to achieve stably 100% of accuracy with $n = 20$ and $\alpha_1 = 95\%$. Regarding Framework B, it stably achieves 98.88% by setting $n = 30$. The results for Weizmann datasets are summarised in Table 3.1, which also reports other state-of-the-art results. It is evident that the proposed frameworks are among the best results reported for the Weizmann dataset.

Additionally, Figure 3.14-(a) provides comparisons between the proposed Framework A and the baseline performance related to different n settings. Weizmann datasets, with 10-actions, is used as a test case. It can be seen that the proposed Framework A achieves superior and stable performance than the baseline, by using smaller values of n.

Regarding i3DPost, despite it is a multi-viewpoints dataset, it can also be used for single-viewpoint experiments. Therefore, by passing to the training and testing process only samples from a fixed point of views, single-viewpoint experiments can be performed and obtained results can be averaged. Thus, the following results are given on average over the eight points of view. My implementation of the baseline achieves 97.46% accuracy with $n = 40$. Instead, with only $n = 10$ ($\alpha_1 = 95\%$), the proposed framework achieves 98.24% accuracy. Figure 3.14-(b) shows comparisons between the proposed Framework A and the baseline results in the 8-actions setting for different values of n. It is evident that, in this case, the proposed Framework is able

to outperform the baseline for all considered values of n. Since i3DPost is mainly a multi-viewpoint dataset, it was not possible to find other state-of-the-art approaches which were considering single-viewpoint experiments based on this dataset. Therefore, the comparison with the state-of-the-art is not possible.

Method	L	Accuracy	n	α_1
Proposed Framework A	10	**100%**	20	95%
Proposed Framework B	10	98.88%	30	99%
Baseline	10	98.66%	30	-
Gorelick et al. [6]	10	**100%**	-	-
Jiang et al. [82]	10	**100%**	-	-
C.Li et al. [48]	9	97.53%	-	-
Ahsan et al. [124]	9	97.5%	-	-
Ahsan et al. [124]	10	94.26%	-	-

Table 3.1: Comparisons for Weizmann dataset (LOAO).

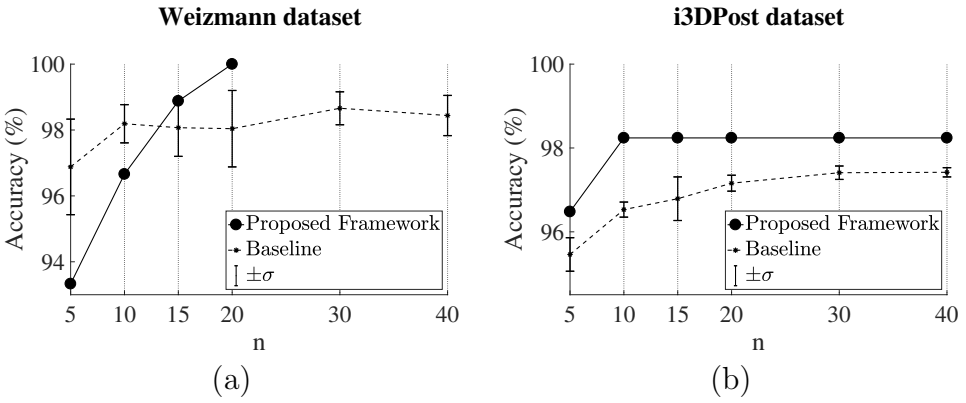


Figure 3.14: Comparison between baseline and proposed pipeline for the small dataset pipeline in Section 3.3.2. In particular, the graphs depict n against accuracy in the LOAO setting, with $\alpha_1 = 95\%$. (a) Results for Weizmann dataset with 10 actions. (b) Results for i3DPost dataset with 8 actions. In this case, only single viewpoint setting is considered. Thus, results are reported on average over the 8 viewpoints included in i3DPost. The proposed Framework saturates at $n = 10$, suggesting that the remaining incorrectly classified action samples contains body movements that 3D-HOG fails in discriminating.

3.4.3 i3DPost (Multi-Viewpoints) Results

In this section, multi-viewpoints experiments are performed to compare the proposed frameworks results with the baseline and other state-of-the-art approaches in a more challenging scenario.

8-actions/8-viewpoints

In this sub section, all actions are considered, including the two multi-targets actions. My implementation of the baseline achieves 98.86% accuracy by setting $n = 30$, with a standard deviation $\sigma = 0.17$. However, the proposed Framework B achieves higher and stable accuracy of 99.60% by setting $n = 30$ and $\alpha_2 = 95\%$. As opposed, the proposed Framework A only reaches 98.04%.

Comparisons with other state-of-the-art results on the i3DPost dataset under the LOAO setting can be rigorously performed as long as the same pre-processing steps are considered, such as ROI detection and background subtraction. However, no standard evaluation protocol has been fixed for this dataset in the literature. Nevertheless, following the comparison suggested by Hilsenbeck et al. [125], Table 3.2 reports the best results reported in the literature with LOAO setting to the best of our knowledge. Methods, where the pre-processing steps are entirely entrusted to the machine, are also highlighted. As can be seen from Table 3.2, the proposed Framework B outperforms the baseline and all other reported methods.

6-actions/8-viewpoints

In this sub section, the two multi-targets actions were removed and only the single-target actions were considered. As summarised in Table 3.2, the proposed Framework B achieved 99.73% accuracy, with $n = 30$ and $\alpha_2 = 99\%$, outperforming the baseline and all other reported methods.

Method	L	W	Accuracy	n	α_2
Proposed Framework B ⊛	8	8	**99.60%**	30	95%
Proposed Framework B ⊛	6	8	**99.73%**	30	99%
Proposed Framework A ⊛	8	8	98.04%	30	99%
Proposed Framework A ⊛	6	8	99.47%	30	99%
Baseline ⊛	8	8	98.82%	30	-
Baseline ⊛	6	8	99.47%	30	-
Castro et al. [126] ⊙	6	2	99.00%	-	-
Iosifidis et al. [57]	6	8	98.16%	-	-
Iosifidis et al. [57]	8	8	96.34%	-	-
Azary et al. [58]	6	8	92.97%	-	-
Hilsenbeck et al. [125] ⊙⊛	6	8	92.42%	-	-

Table 3.2: Accuracy results for i3DPost dataset (LOAO). The ⊙ highlights methods with automatic selection of ROIs, while ⊛ highlights methods with automatic background subtraction without prior knowledge.

3.4.4 Computational Cost and Complexity

In this section, the computational complexity of the most expensive step of the methods discussed in this chapter is firstly discussed. Moreover, the execution times for the baseline, the Framework A and B are compared.

In the baseline and the proposed frameworks, the embedding procedure is the most expensive step in terms of computational effort. In fact, it depends on the time-length K of the samples, the number of actions L and on the size n of the library V. Moreover, it is required in both the training and testing phases. For this reason, only this step computational complexity is analysed in this section.

For each location p, the embedding vector $\mathcal{D}_p$ in (3.2.2) is composed of nLW entries, each of them computed in (3.2.1) as a minimum among one-to-many comparisons. In formulas, the complexity of the embedding procedure can be expressed with respect to n as

$$f(n) = (c_1 + Kc_2 + \mathcal{O}(K))nLW = \mathcal{O}(n) \tag{3.4.1}$$

where c_1 and c_2 are positive constants and where it is assumed that the complexity for the minimisation over array problem in (3.2.1) is, in the worst case, $\mathcal{O}(K)$. Clearly, high number of actions L, high number of viewpoints W and high time-length K worsen performance. However, L, W and K are structural parameters which are fixed for the considered dataset. The only parameter which can be potentially optimised is n. Therefore, equation (3.4.1) shows that choosing n as low as possible is important for fast-computation applications. Results in Sections 3.4.2 and 3.4.3 show that the proposed frameworks achieve better performance in terms of accuracy for equivalent n than the baseline. Thus, proposed frameworks are preferable to the baseline in computationally-efficient implementations.

In Figure 3.15, the testing phase the execution time for the baseline,

Framework A and B is reported. It can be seen that the three methods have the same execution time for the first three steps, i.e. background subtraction, 3D-HOG and embedding. This is due to the fact that: 1) the background subtraction processing is the same for all methods; 2) the 3D-HOG computation is performed with the same parameters, to allow results comparisons; 3) the embedding step is performed by setting $n = 30$ in all cases. However, it can be seen that the baseline and Framework A last step, i.e. classification, is considerably slower than the Framework B classification step. This is due to the fact that the baseline and Framework A perform 48 local classifications. In contrast, Framework B replaces the 48 classifications with 48 matrix-to-vector multiplications (to perform the PCA transformation computed in the training phase), and performs a single classification step. Overall, since the real-time threshold for video-based processing can be reasonably set to 19 FPS, both the baseline and the proposed approaches do not perform within real-time performance.

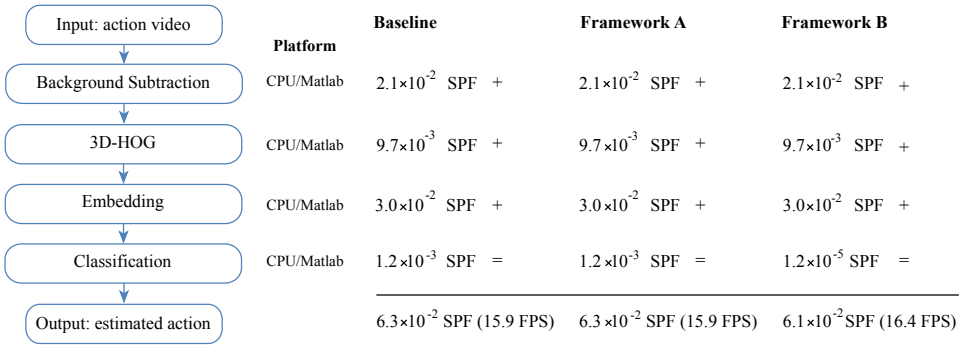


Figure 3.15: Execution time comparison for the baseline, Frameworks A and B. Each processing component execution time is independently measured in Seconds-Per-Frame (SPF). The cumulative SPF is computed for each method, and transformed in Frame-Per-Second (FPS). There results have been obtained by Matlab code implementations running on a Windows 7 workstation, 64-bit, with a CPU Intel Core i5-6600 @ 3.30GHz 3.30GHz, 16 GB of RAM, running Matlab implementations.

3.4.5 Time Windows Examples

In this section, a representative sample from i3DPost dataset is considered to study how Framework B performance changes according to the time window position. The sample consists of 56 frames showing two targets performing a *pull* action. The first half of the clip consists of action preparation, i.e. the two targets join their hands. In the second half of the clip, one of the two targets actually starts pulling the other target.

In Figures 3.16 to 3.20, the Framework B output is shown for different time windows. The considered time windows increase in length from 1st to 8th frames up to the 1st to 40th frame. The reported output is always, mistakenly, *hands-shake*. This is due because the performed action in the considered time windows strongly resemble shaking hands.

The output changes when the time window starts including the most representative part of the pulling action, i.e. from 40th to 58th frame. In Figures 3.21 to 3.27, it is shown that Framework B correctly classify the pulling action when the considered time window includes the most representative action part.

It is interesting to notice that, despite the considered viewpoint induced one target's body to occlude the other target's body, no viewpoint-related issues on the classifications are noticeable.

This study confirms what it was reasonable to expect. When the pulling action is actually covered by the considered time window, the classification is performed correctly. If the time window does not adequately cover the pulling action, there is a high chance of misclassification.

3.5 Critical Analysis

The purpose of this section is to critically analyse the proposed methods, and evaluate their performance with respect to the aims and objectives set

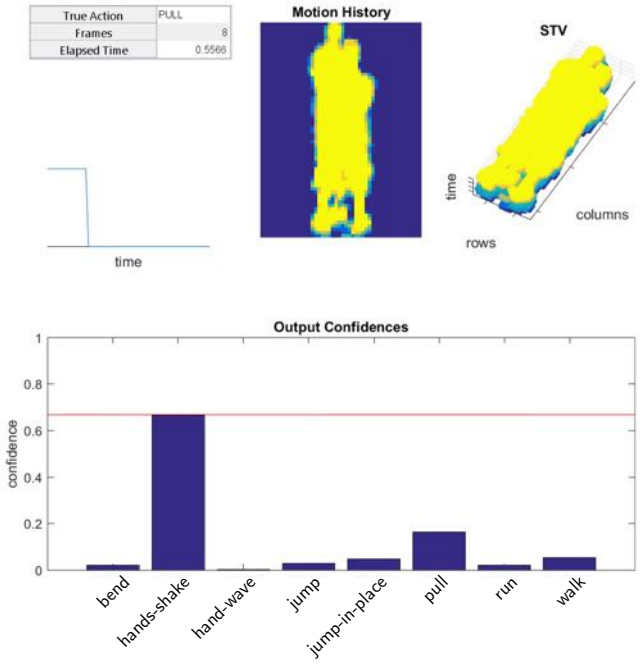


Figure 3.16: *pull* action (frames 1-8). Framework B mistakenly outputs *hand-shake* with high confidence. This is due to the fact that, so far, the performed action looks like an hand shaking.

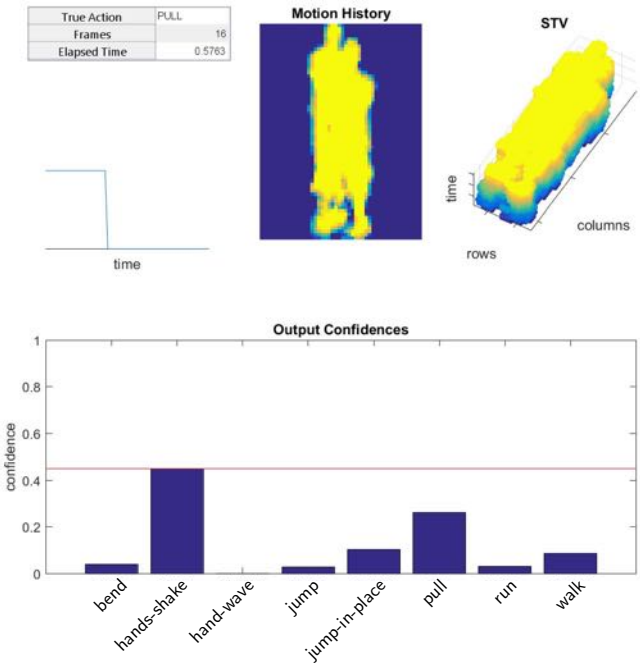


Figure 3.17: *pull* action (frames 1-16). Framework B mistakenly outputs *hand-shake*. However, additional 8 frames made the previous confidence to drop. This is due because the performed action started look slightly different than an hand shaking.

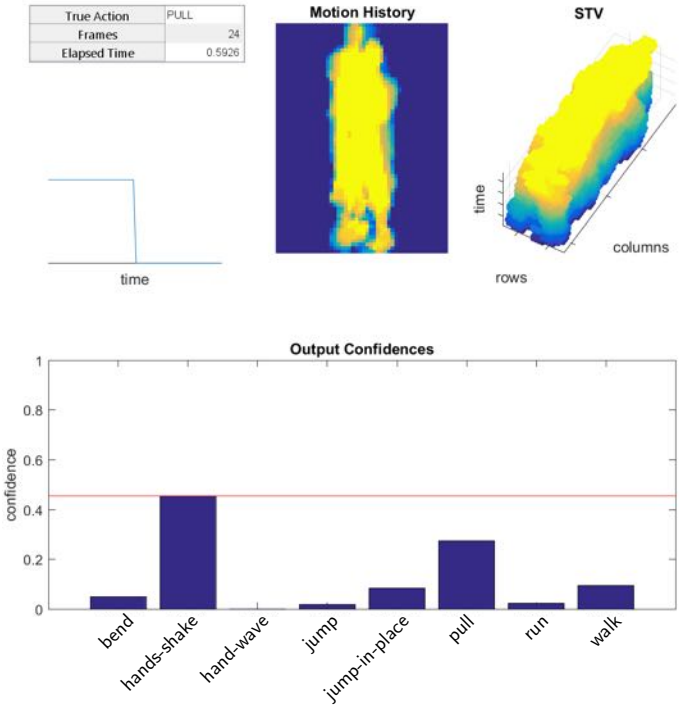


Figure 3.18: *pull* action (frames 1-24). Framework B mistakenly outputs *hand-shake*. The confidence is still the same than the previous case.

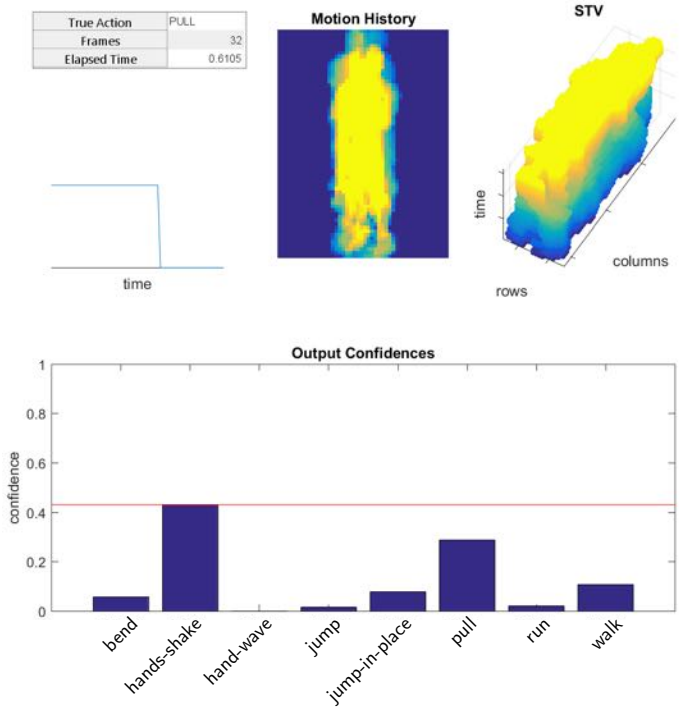


Figure 3.19: *pull* action (frames 1-32). Framework B outputs *hand-shake* . However, the confidence of *pull* is getting closer to the confidence of *hand-shake*.

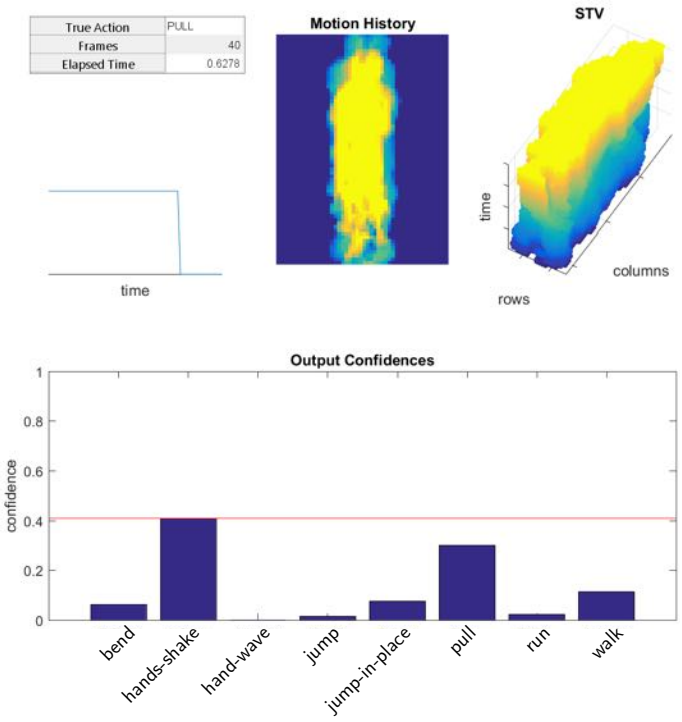


Figure 3.20: *pull* action (frames 1-40). Framework B still outputs *hand-shake* . However, the confidence of *pull* is similar.

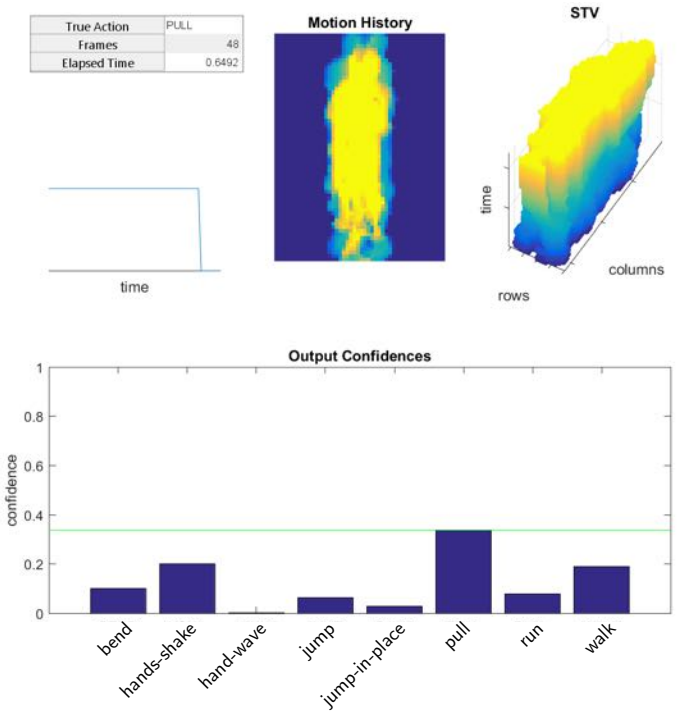


Figure 3.21: *pull* action (frames 1-48). Framework B correctly outputs *pull*. However, the output confidence is low.

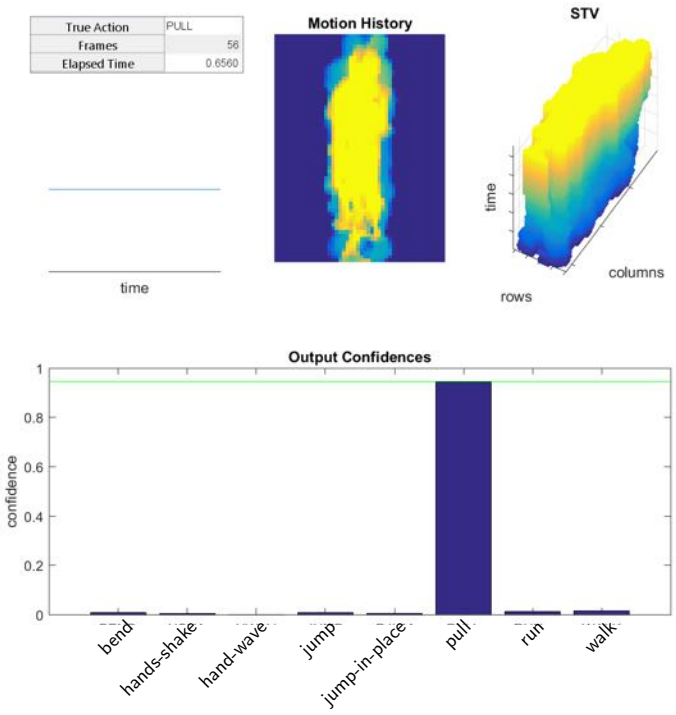


Figure 3.22: *pull* action (frames 1-56). Framework B correctly outputs *pull* with high confidence.

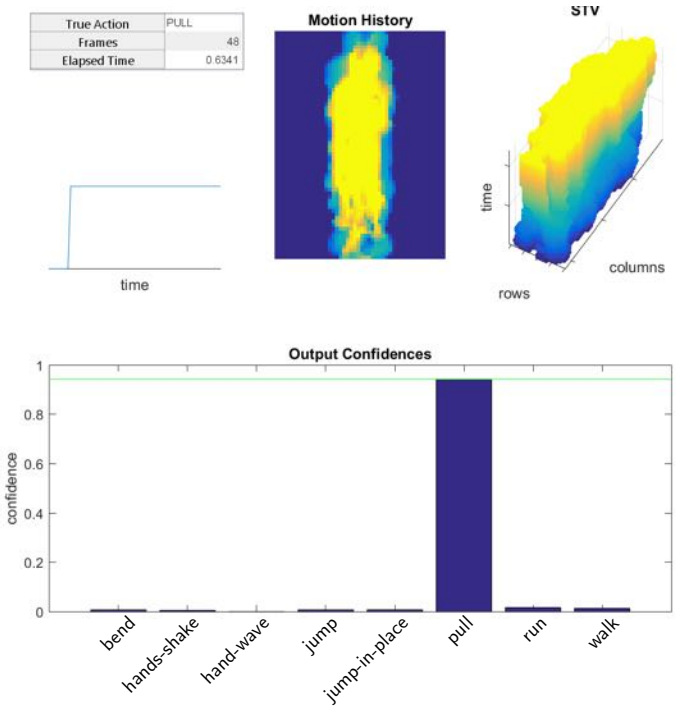


Figure 3.23: *pull* action (frames 9-56). Framework B correctly outputs *pull* with high confidence.

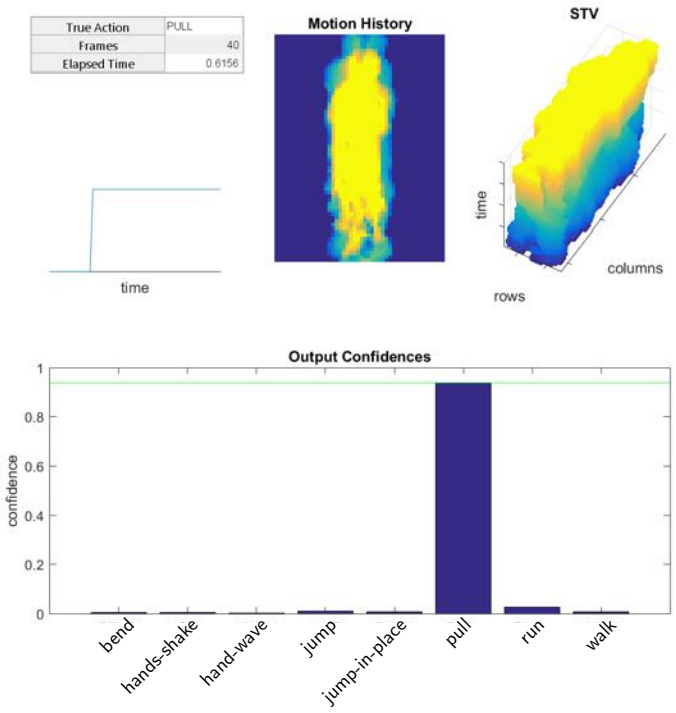


Figure 3.24: *pull* action (frames 17-56). Framework B correctly outputs *pull* with high confidence.

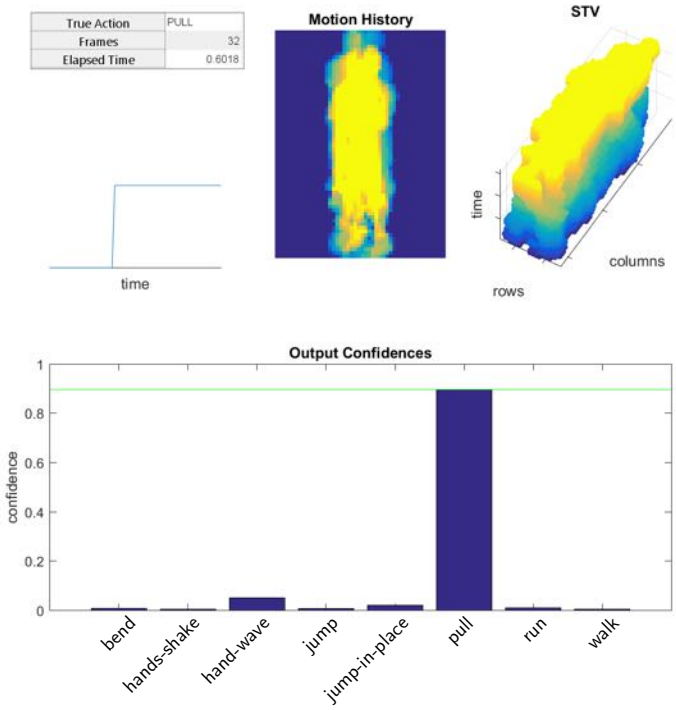


Figure 3.25: *pull* action (frames 25-56). Framework B correctly outputs *pull* with high confidence.

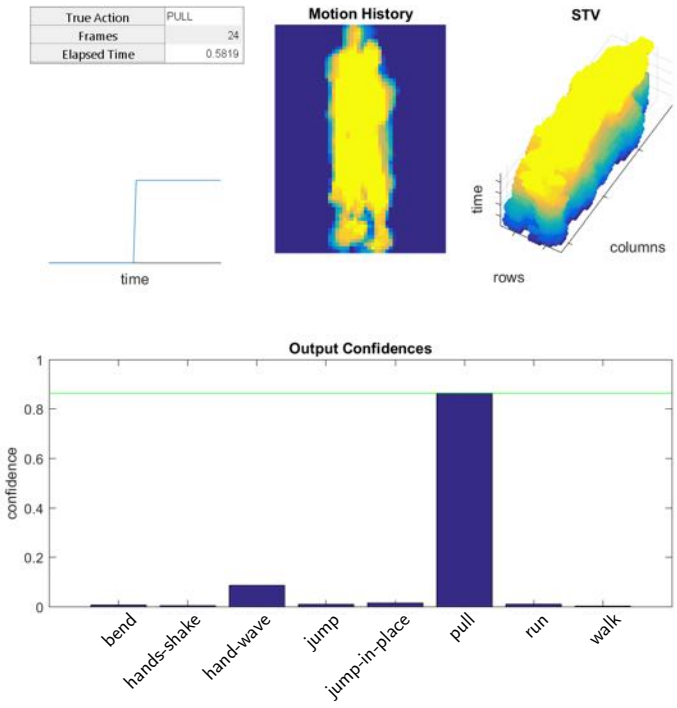


Figure 3.26: *pull* action (frames 33-56). Framework B correctly outputs *pull* with high confidence.

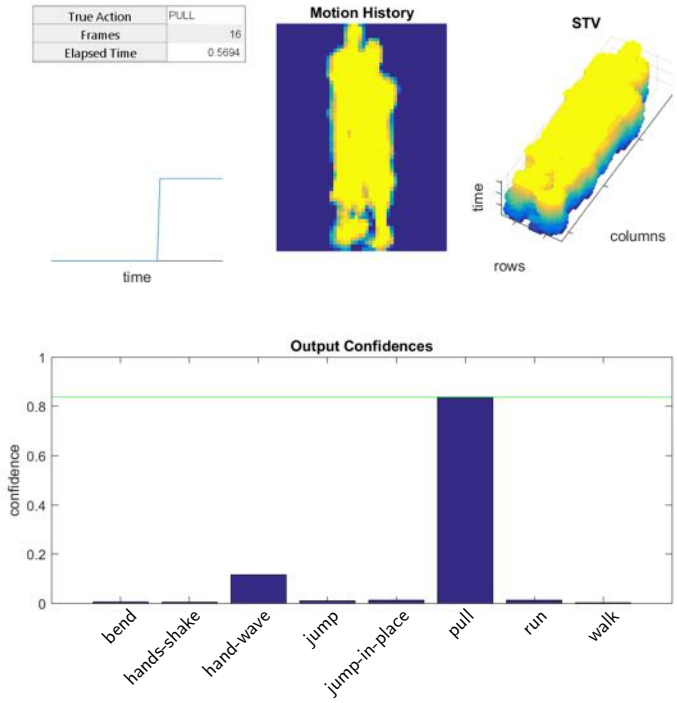


Figure 3.27: *pull* action (frames 41-56). Framework B correctly outputs *pull* with high confidence.

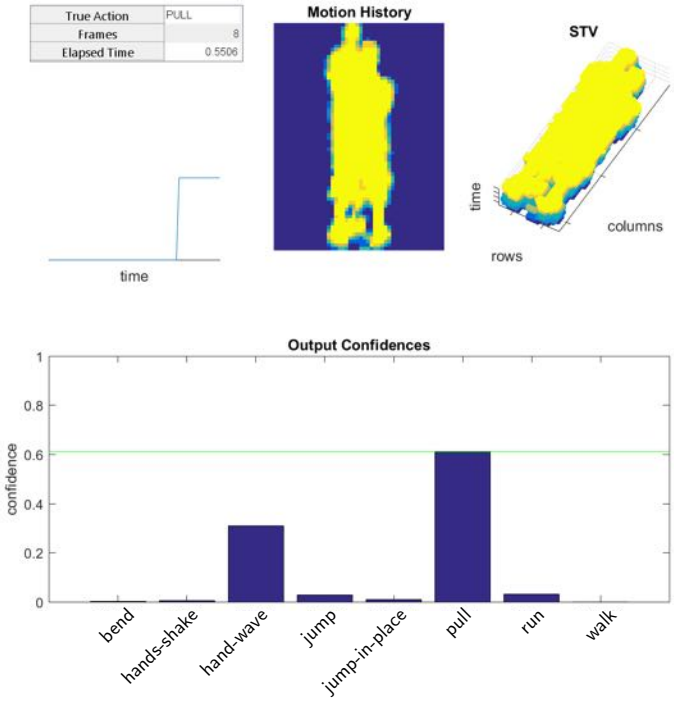


Figure 3.28: *pull* action (frames 48-56). Framework B correctly outputs *pull* with high confidence. However, the *hand-shake* confidence increased.

in Section 1.3.

Objective 1 is achieved, since the proposed frameworks outperform the state-of-the-art on the tested datasets. In particular, this chapter studies made evident that, although the Framework B is a simplification of Framework A, it seems to perform better on datasets with a considerably high number of samples, i.e. i3DPost. This is probably due to the fact that the PCA step is more reliable when training data is sufficiently numerous. However, further studies on other datasets would be required to confirm this statement.

Objective 2 is achieved, since the proposed frameworks are based on a non-random process, which ensures stability over different training rounds.

Overall, the promising results obtained by the proposed Frameworks A and B rely on a more sophisticated method to manage flows information. The baseline method is based only on flows and action labels l, on the assumption that each flow carries the same amount of information. However, within a certain flow at location p, not all the actions are distinguishable, implying mistakes and misclassifications in the learning stage. This drawback motivated the proposed frameworks development.

Objective 3 is addressed in this section. On the basis of the above-mentioned achievements, the reliability for real-world problems was assessed by conducting a conclusive set of experiments on KTH, IXMAS and new video dataset, named ISLD-2017, recorded in the Intelligent Sensing Lab. ISLD-2017 consists of multi-targets video recordings, where the targets perform simple posture based actions such as walking, standing, sitting, opening doors and pulling. The purpose of ISLD-2017 was only to serve as a preliminary working platform for future recordings and to assess the reliability of silhouette based methods in a non-cooperative scenario. In particular, the goal was to identify additional challenges that real-world applications propose, and that tested datasets did not highlight. In Figures 3.29 and 3.30,

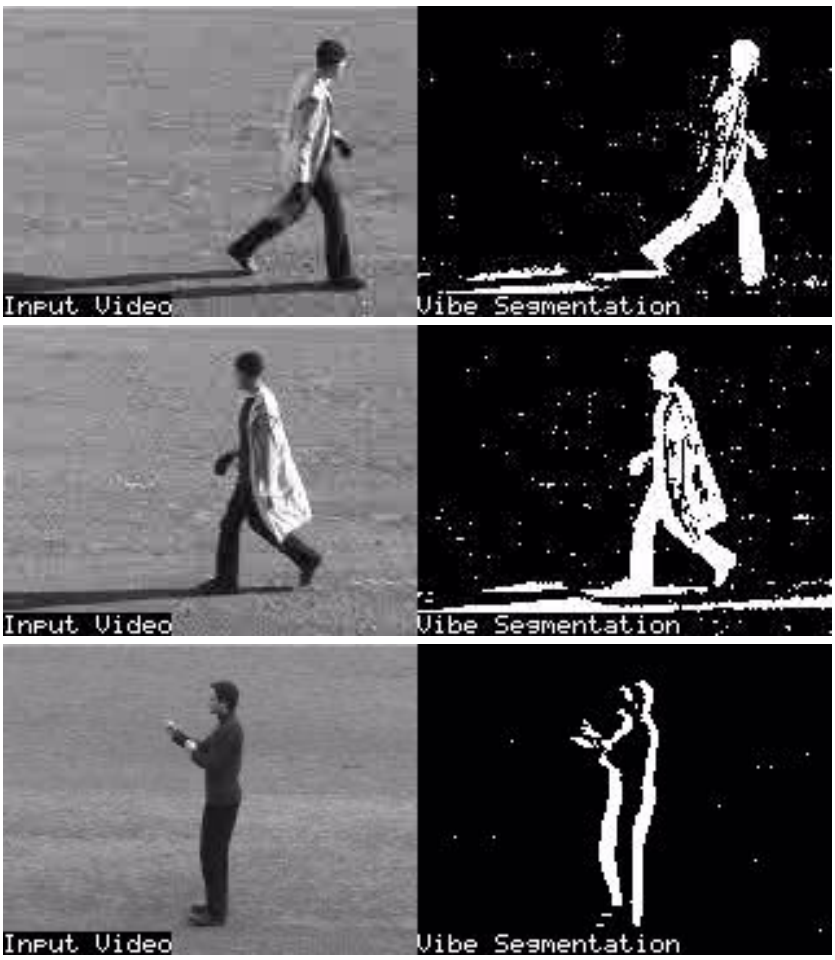


Figure 3.29: Frames from KTH Dataset [7] processed with ViBe for background subtraction. *Ghost* and shadows effects are visible, which strongly corrupt the detected silhouettes.

three background subtraction masks obtained from the KTH and IXMAS datasets are provided. In Figures 3.31 to 3.35, five background subtraction masks obtained from ISLD-2017 are shown.

The examples shown in this section reveal the the following unexpected, real-world related issues that were encountered:

- Despite ViBE is claimed as a state-of-the-art method for background subtraction, it is far from being as robust as required. In particular, *ghost* effects [39] strongly affect the final output, despite the intense ViBE hyper-parameters selection performed to compensate for this

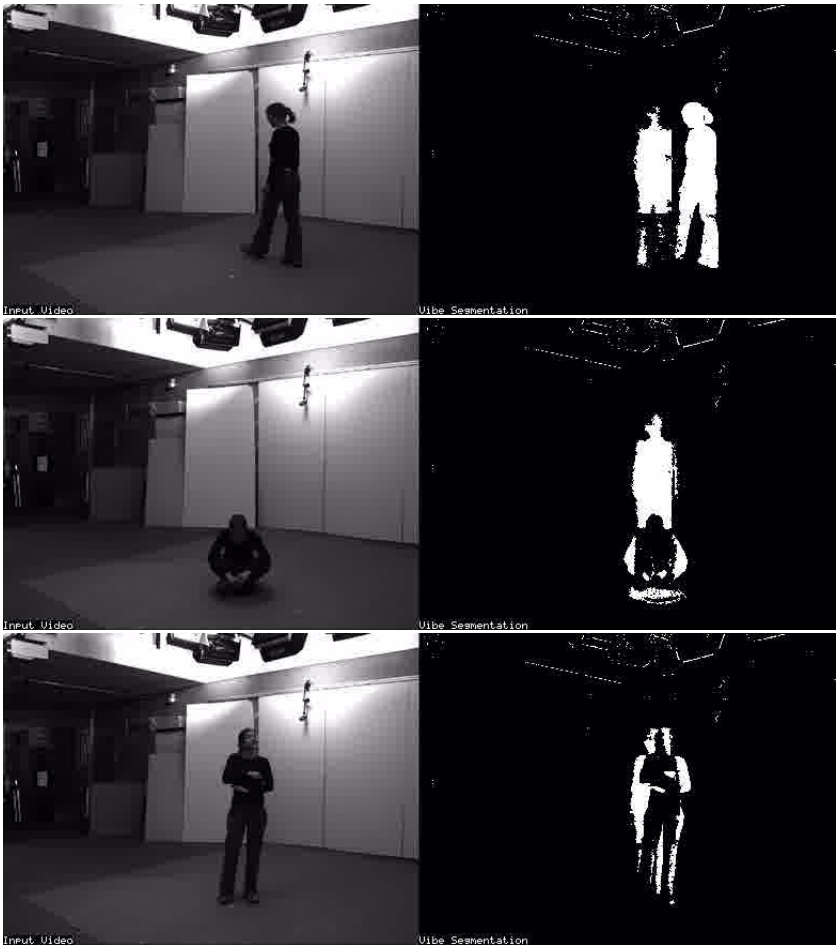


Figure 3.30: Frames from IXMAS Dataset [9] processed with ViBe for background subtraction. *Ghost* effects are visible, which strongly corrupt the detected silhouettes.

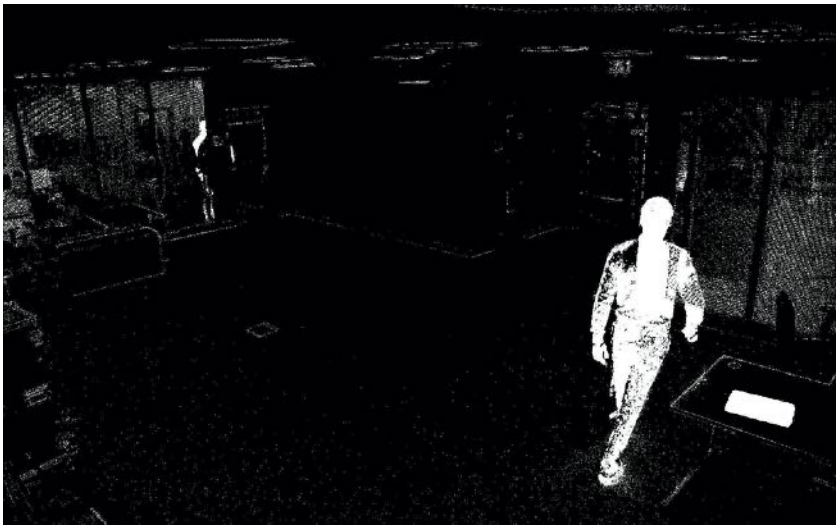


Figure 3.31: Frame from ISLD-2017 Dataset processed with ViBe for background subtraction. In the right bottom corner, it is visible the shape of a laptop. Therefore, a detector to distinguish the target from the laptop would be required as further processing step.

Figure 3.32: Frame from ISLD-2017 Dataset processed with ViBe for background subtraction. Since two targets are present in the scene, a detector would be required to distinguish from each other, other than distinguish them from other visible shapes, e.g. the chairs, the doors and the laptop.

Figure 3.33: Frame from ISLD-2017 Dataset processed with ViBe for background subtraction. Two targets are clearly visible, while a third sitting target shape is disturbed by the chairs shape. A *ghost* effect is also visible behind the door on the left upper corner. Therefore, a detector would be required to help in distinguish the targets data from other disturbing factors.

Figure 3.34: Frame from ISLD-2017 Dataset processed with ViBe for background subtraction. Two targets are clearly visible, while a third target is moving from the sitting to the standing position. This movement produces artefacts which heavily compromises data clarity. In fact, the third target shape is barely recognisable.

Figure 3.35: Frame from ISLD-2017 Dataset processed with ViBe for background subtraction. Three targets are clearly visible since no other nearby potentially moving objects are visible. However, a detector would be required in any case, since the three targets data needs to be distinguished from each other and from other objects, including the human *ghost* persistently visible on the left upper corner.

problem.

- It seems difficult to obtain homogenous silhouettes from different datasets. Due to the above-mentioned ViBE-related issues, since different datasets have different light and background conditions, the obtained mask are qualitatively dissimilar to each other. This issue suggests that a silhouette-based method is hard to be generalised or to be tested in cross-dataset settings, unless plenty of data is available for training.

- Background subtraction methods are sensitive to any moving objects, regardless of their semantic. Therefore, it turns out that multiple objects shapes and humans silhouettes can simultaneously be present on the same mask. Therefore, additional processing is required to perform *detection*, to properly distinguish the human targets from other objects and to draw the target ROIs.

Therefore, it is reasonable to expect that additional computational band-

width would be required to further pre-process RGB data or silhouettes, to overcome the above-mentioned issues. However, as discussed in Section 3.4.4, the processing time that it has been possible to achieve in these simulations do not seems to leave much room for further processing. In other words, it seems reasonable to argue that silhouette-based HAR, in the form discussed in this Figure 3.15, is slowed down by mainly the background subtraction method and the embedding step. Therefore, the silhouette-based HAR, as studied in this thesis, does not seem to be a promising direction for computationally-efficient implementations.

To overcome the above-mentioned problems, it might be more computationally efficient to prefer another input data for HAR, in place of human silhouettes, for computationally-efficient applications. Other possible alternatives are the following:

- Deep learning-based human pose detectors such as OpenPose can, jointly, provide human detection and human poses. Therefore, the above-mentioned issues would be solved if human poses would be a reliable source of information for HAR, compared to human silhouettes. This option is investigated in Chapter 4.

- Deep learning-based human detectors, such as YoLo [11] or SSD [10] can effectively and efficiently draw a bounding box around the human target, frame-by-frame, at the RGB level. Therefore, the human RGB-based ROI can be provided as input for subsequent HAR processing. This option is investigated in Chapter 5.

3.6 Chapter Summary

In this chapter, Objectives 1-3 of this thesis have been achieved. In particular, a silhouette-based method for HAR has been replicated and implemented as a baseline. Therefore, two frameworks for silhouette-based HAR

have been proposed. As extensive experiments have revealed, both proposed methods outperform the baseline and the state-of-the-art on the tested datasets for silhouette-based HAR. The results showed that the advantages of the proposed frameworks are in terms of accuracy and performance stability over different training rounds. A critical analysis of the proposed frameworks was also performed, based on the results achieved on publicly available datasets and new data recorded in the Intelligent Sensing Lab. The conclusion was that despite the proposed methods are effective for HAR, they might be not the best approach for computationally-efficient implementations. This conclusion motivates the further conducted studies, which are the objective of the following chapters.

Chapter 4

2D POSE-BASED REAL-TIME HUMAN ACTION RECOGNITION WITH OCCLUSION-HANDLING

4.1 Introduction

In Chapter 3, the accuracy results obtained by the proposed silhouette-based approaches were encouraging. However, the experiments conducted on the ISLD-2018 dataset highlighted limitations regarding computational efficiency and practical implementation in non-cooperative scenarios.

In this chapter, the idea is to overcome silhouette related limitations discussed in Chapter 3 by exploiting 2D human-poses provided by a popular detector, i.e. OpenPose [12]. As already discussed in Chapter 2, human-poses carry, frame-by-frame, body posture-related information in the form of 2D landmarks. Moreover, human-poses are explicitly providing human-target detection. Thus, by using human-poses, the detection step, which was missing in the proposed methods in Chapter 3, and the posture-based information retrieval are fused into the same step. Additionally, the human target detections allows easy-to-implement tracking for multiple-target test-

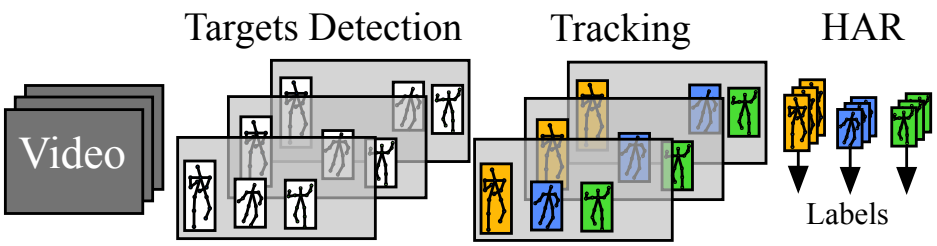


Figure 4.1: Overall HAR framework. CCTV-like videos are pre-processed with a human pose detector, such as OpenPose, to estimate the positions of the targets (bounding box) and their body limb positions (landmarks). Subsequently, tracking algorithms provide targets identities, allowing consistent grouping of detected targets data for HAR. Finally, for each detected target, the action label is estimated by using the tracked data.

ing. For these reasons, 2D human-poses are investigated in this chapter as a new source of information for HAR. It is important to remark that the results shown in this chapter represent one of the first, few, simultaneous attempts to perform HAR by using 2D human-poses. This makes the contributions of this chapter of breakthrough importance in the literature.

This chapter contributions are based on the detection and tracking shown in Figure 4.1, which has been already explored for tracking-related problems [84–86, 127, 128]. In the case of multi-target analysis based on CCTV-like recordings, the first step generally consists of the detection of human targets by using RGB data. This step requires significant computational effort. Subsequently, the tracking retrieves the identities of the targets and allow consistent time-wise data association based on targets identities. The tracking relies on detected bounding boxes or body landmarks coordinates rather than on the RGB data. Thus, this step is computationally efficient and each tracked target RGB data can be further processed for HAR.

As shown in this chapter, OpenPose is relatively prone to false-detections due to target occlusions or cluttered data. This problem has been already explored and documented in a recent work for tracking applications [84]. Thus, this chapter also focuses on new HAR strategies which are robust to occlusion and missing data problems. The proposed algorithm is named

ActionXPose, and it is based on robust handcrafted features extracted from landmark time-sequences. The proposed features are further processed by Multivariate Long Short-Time Memory and Fully Convolutional Network (MLSTM-FCN) for classification.

Moreover, this chapter proposes a new dataset, namely the Intelligent Sensing Lab Dataset (ISLD), which is specifically designed for posture-based HAR. However, existing datasets are also explored to perform extensive evaluation. To evaluate ActionXPose performance, single-dataset, multi-datasets and cross-datasets testings are performed. Moreover, this chapter includes the ablation study, the occlusions study and the video quality study. Results show that ActionXPose outperforms existing methods in most of the tests and shows greater robustness to occlusion and missing data problems.

In conclusion, the main contribution of this chapter is threefold: 1) a new, real-time algorithm for posed-based HAR; 2) new occlusion-handling techniques for robust pose-based HAR; 3) a new dataset, named ISLD, for pose-based HAR. These contributions were obtained by addressing Objectives 4-6 introduced in Section 1.3.

The rest of this chapter is organised as follows. In Section 4.2, the baselines and the ActionXPose algorithm are introduced, with emphasis on the proposed low-level features extraction, the proposed high-level features computation and proposed occlusion-handling strategies. In Section 4.4, several experiments are presented to extensively evaluate the performance of ActionXPose. Finally, in Section 4.5, conclusions are drawn, and the links with the next chapter are provided.

4.2 Methodology

4.2.1 Baseline Methods

In this section, three baseline methods are defined. The baselines are defined by borrowing techniques from the 3D skeleton-based HAR field, due to its similarities with the proposed 2D pose-based HAR.

The sequence $p_i(t)$ provided by OpenPose are location and body size dependent [69]. Thus, translation and scaling are required in order to normalise data across different samples. Specifically, the location dependency problem can be addressed by transforming $p_i(t)$ from the *absolute* to the *root-centred* coordinate reference system. The transformed pose coordinates are defined as:

$$(\bar{x}_j, \bar{y}_j)_i = (x_j, y_j)_i - (x_2, y_2)_i \quad \forall j \in J, \tag{4.2.1}$$

where the $\bar{(\cdot)}$ operator denotes the centring transformation and where the dependence of t has been conveniently omitted. Thus,

$$\bar{p}_i = \{(\bar{x}_j, \bar{y}_j)_i\}_{j \in J}, \tag{4.2.2}$$

is the set of root-centred coordinates defined by (4.2.1).

Furthermore, let $\bar{\bar{(\cdot)}}$ be the scaling operator, defined as follows:

$$\bar{\bar{p}}_i = \{(\bar{\bar{x}}_j, \bar{\bar{y}}_j)_i\}_{j \in J} \tag{4.2.3}$$

where $\bar{\bar{p}}_i$ is obtained by scaling $\bar{p}_i(t)$ coordinates by using the following constraint

$$\bar{\bar{v}}_{j_1,j_2} = \frac{\bar{v}_{j_1,j_2}}{\|\bar{v}_{2,9}\|_2} \quad \forall j_1, j_2 \in J \tag{4.2.4}$$

where $\bar{v}_{2,9}$ is the vector link between the root and the right hip landmarks and $\|*\|_2$ is the Euclidean norm operator. Due to (4.2.1) and (4.2.4), the

target position and the size information are discarded.

According to the proposed definition for $p_i(t)$, these sequences mostly contain *spatial* information about the motion. To obtain *temporal* information, $p'_i(t) = p_i(t+1) - p_i(t)$ can be defined. For the rest of this chapter, $p_i(t)$ denotes the transformed poses in (4.2.3).

Inspired by existing literature related to 3D skeleton-based HAR, three baseline methods has been defined as follows.

Baseline A: [69] + [129]

Baseline A consists of a simple learning step based on normalised OpenPose coordinates sequences $p_i(t)$ and $p'_i(t)$. A Multivariate LSTM-FCN architecture with a *time-based attention mechanism* (MLSTM-FCN) [129] is used for the classification step. This algorithm takes as input the coordinate sequences obtained from training data $\mathbb{T}$, including action labels l_i, to train a supervised classification model to be tested on $\mathbb{T}^*$. MSLTM-FCN has been chosen as a time sequences-based classification method due to its state-of-the-art performance on several datasets for different problems including HAR.

Baseline B: [68] + [129]

Baseline B consists of computing mutual OpenPose landmarks distances [68] and exploiting the obtained time-sequences for classification. The classification step is again performed by using the above-mentioned MLSMT-FCN architecture.

Baseline C: [68] + [69] + [129]

Baseline C consists of a hybrid approach obtained by merging the previous two baselines. Thus, $p_i(t)$, $p'_i(t)$ and mutual distances between landmarks are considered for classification. The classification step is again performed

by using the MLSMT-FCN architecture.

Formally, $p_i(t)$ and $p'_i(t)$ represent *low-level* features, which can be provided to the MLSTM-FCN for classification. In particular, by using $p_i(t)$ and $p'_i(t)$, the MLSTM-FCN performs landmark-based attention due to its architecture. However, such levels of detail can be confusing in some cases, due to high intra-class similarities or within-class variations. Moreover, when some landmarks are persistently missing due to occlusions, the corresponding sequences will be completely lost, compromising the robustness against unexpected occlusions. For example, in the cases of Baseline B and C, a single, persistent missing landmark $(x_j, y_j)_i$ not only neglects two x_j and y_j sequences, but also compromises the calculation of mutual distances.

Therefore, the next section provide novel high-level features, that are designed to be robust to missing data, and additional occlusion-handling methods, providing an effective solution to this problem.

4.3 Proposed ActionXPose

4.3.1 Defining Poses Libraries

The main goal of this section is to exploit training data $\mathbb{T}$ to learn general poses that best represent each action, from each viewpoint. In other words, the output of this step is a *pose library*.

Since root coordinates in $p_i(t)$ were set to zero in the previous section, let $u_i(t) = (x_1, y_1, x_3, y_3, \ldots, x_J, y_J) \in \mathbb{R}^{2J-2}$ be the vector obtained by unrolling $p_i(t)$ and skipping the root coordinates $(x_2, y_2)_i$. The unsupervised clustering method Self-Organizing Map (SOM) [21] is used in a semi-supervised fashion, to explore natural clusters in $\mathbb{R}^{2J-2}$. Since the SOM algorithm expects no missing data, in this stage, body left/right symmetry was exploited for dealing with possible persistent occlusions occurred in training

data, mostly due to target self-occlusions. In particular, persistently missing landmarks were estimated by mirroring available data. For example, if the left-shoulder was missing, data was filled with the transformed right-shoulder obtained by mirroring it with respect to the root landmark.

Additionally, SOM requires a cluster topology to be defined. Since prior-information about the distribution of pose data is not provided, the homogeneous topology, $[q, \dots, q] \in \mathbb{R}^m$, is set for a given integer q and a given space dimension m. This choice forces the SOM architecture to have q^m neurons linked each other with a homogenous rectangular topology, defining q^m clusters. Since the SOM computational time is affected by either the number of considered vectors $u_i(t)$, the q topology parameter and the space dimension $2J-2$, a trade-off between these parameters is required.

To solve this problem, let $\tilde{u}_i(t)$ be the vector containing the first m principal components of $u_i(t)$ obtained through the PCA, i.e. $\tilde{u}_i(t) \in \mathbb{R}^m$. Simulations suggest that the best values for m and q are $m = 3$ and $q = 4$, which balance the SOM computational cost while producing a reasonable number of prototypes. In Figure 4.2, comparisons of SOM computational times are provided. Therefore, the whole process can be summarised as follows

$$\mathbb{R}^{2J-2} \xrightarrow{\text{PCA}} \mathbb{R}^m \xrightarrow{\text{SOM}} q^m \text{ clusters} \tag{4.3.1}$$

Thus, for a fixed action label l and a fixed point of view w, the SOM is trained over

$$\{\tilde{u}_i(t) \mid l_i = l, w_i = w\} \subset \mathbb{T}. \tag{4.3.2}$$

This provides an additional cluster label k_i for each training pose $\tilde{u}_i(t)$, as follows:

$$\begin{aligned} \{\tilde{u}_i(t) \mid l_i = l, w_i = w\} &\xrightarrow{\text{SOM}} \{\tilde{u}_i(t) \mid l_i = l, w_i = w, k_i = k\} \\ \forall\, l \in \mathcal{L}, \quad \forall\, w \in \mathcal{W}, \quad k \in \{1, \dots, q^m\}. \end{aligned} \tag{4.3.3}$$

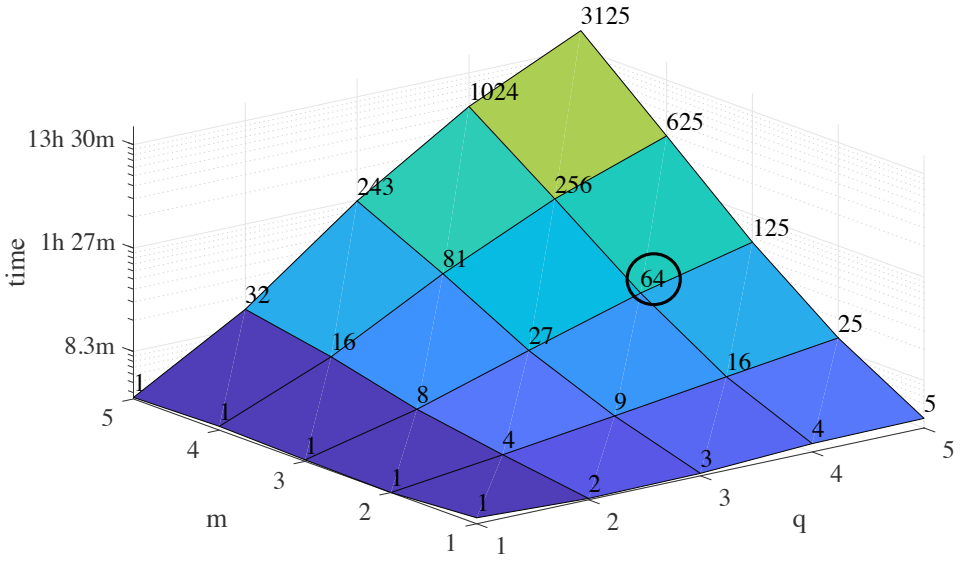


Figure 4.2: SOM time computation comparison. Different values for m and q have been set for the SOM computation. Reported times refer to entire library creation process for the MPOSE dataset. In the graph, q^m values are also reported. As shown, $q^m = 4^3 = 64$ is chosen as trade-off between computation time and number of prototypes in the libraries.

Thus, q^m pose prototypes are defined by averaging cluster labels k_i as follows:

$$U_{l,w,k} = \frac{1}{n_k} \sum_{i,t} \{u_i(t) \mid l_i = l, w_i = w, k_i = k\}$$
$$\forall\ l \in \mathcal{L}, \quad \forall\ w \in \mathcal{W}, \quad k \in \{1, \ldots, q^m\}, \tag{4.3.4}$$

where n_k represents the number of poses within cluster k. In conclusion of this step, the libraries of prototypes are collected from training data as follows:

$$V_l = \left\{ \{U_{l,1,k}\}_{k=1}^{q^m}, \ldots, \{U_{l,|\mathcal{W}|,k}\}_{k=1}^{q^m} \right\} \quad \forall\ l \in \mathcal{L}. \tag{4.3.5}$$

Thus, V_l contains pose prototypes in the form of points in a multidimensional space $\mathbb{R}^{2J-2}$, which are able to cover all variation of considered viewpoints. For a visual example of the V_l set, see Figure 4.3.

Libraries for the temporal information can be similarly defined as follows:

$$S_l = \left\{ \{U'_{l,1,k}\}_{k=1}^{q^m}, \ldots, \{U'_{l,\mathcal{W},k}\}_{k=1}^{q^m} \right\} \quad \forall\ l \in \mathcal{L}, \tag{4.3.6}$$

where $U'_{l,w,k}$ represents prototypes obtained by clustering temporal vectors $u'_i(t)$.

4.3.2 Strategies for Occlusion-Handling

Occlusions, self-occlusions or ambiguous RGB data can affect OpenPose performance, resulting in *persistent* or *short-time* missing data. In this section, we propose four complementary strategies to deal with these problems.

High-level Features

In this section, the problem of persistent occlusions is addressed. A persistent occlusion occur when one or more landmarks are missing for the *entire* sequence. To address this problem, the idea is to exploit the Spatio-temporal libraries V_l and S_l for $l \in \mathcal{L}$ defined in the previous section, to generate high-

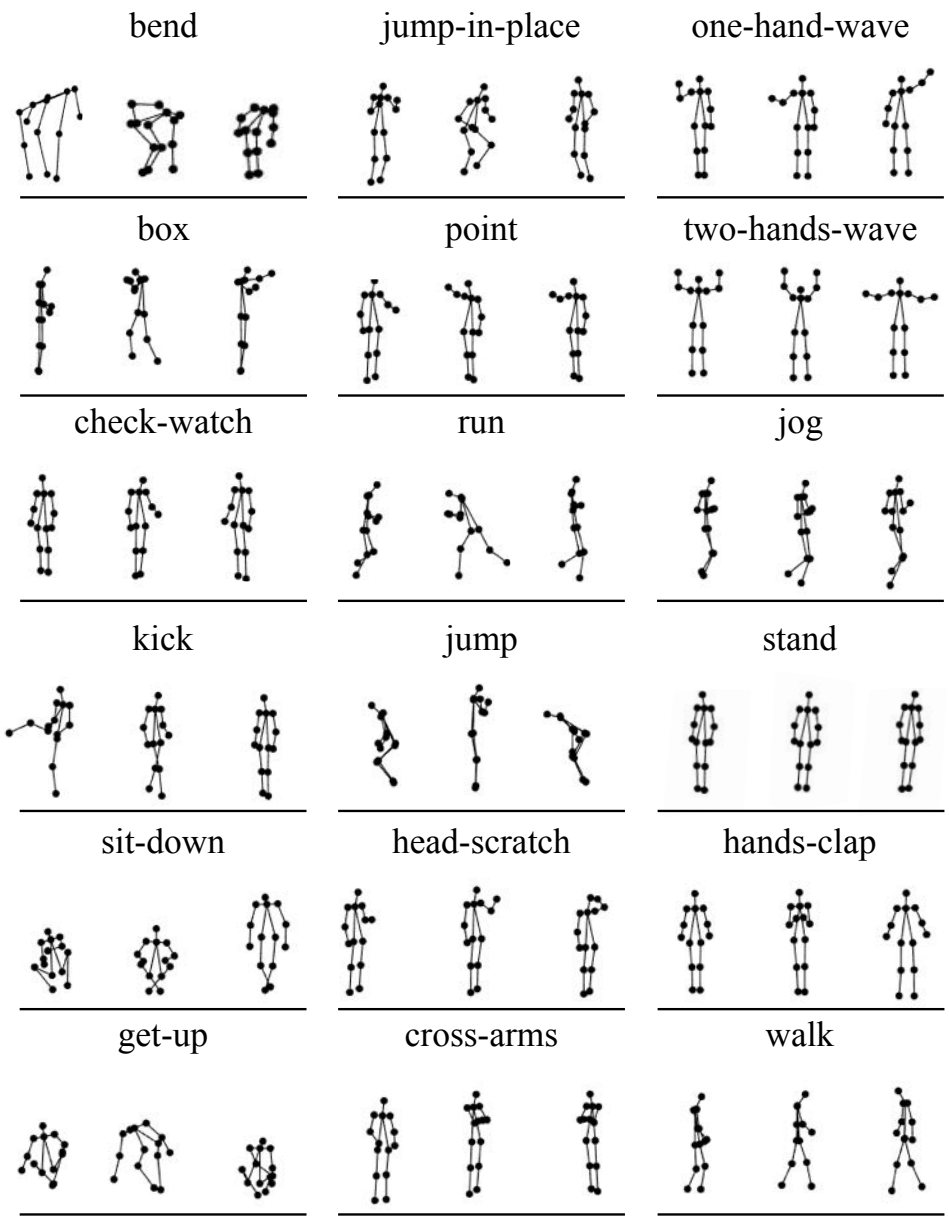


Figure 4.3: Spatial library prototype examples. Three prototypes are randomly selected from V_l, for all $l \in \mathcal{L}$.

level features in the form of time sequences. Inspired by [72], since different body parts carry different information, the idea is to exploit full-body and local-limb attention.

Given $J = \{1, \ldots, 14\}$, let J_a, J_b, J_c, J_d be the landmark subsets as

defined in Figure 1.6; namely,

$$J_a = \{3, 4, 5\} \subset J \qquad J_b = \{6, 7, 8\} \subset J$$
$$J_c = \{9, 10, 11\} \subset J \qquad J_d = \{12, 13, 14\} \subset J. \tag{4.3.7}$$

Let $d_{J_*}(p_i(t), v)$ be the average distance between the generic pose $p_i(t)$ and the generic prototypes $v \in V_l$, computed for landmarks J_*, where J_* represents either J_a, J_b, J_c, J_d or J, and it is defined as follows:

$$d_{J_*(t)}(p_i(t), v) = \frac{1}{|\bar{J}_*(t)|} \sum_{j \in \bar{J}_*} \|(x_j, y_j)_i - (x_j, y_j)_v\|_2 , \tag{4.3.8}$$

where $(x_j, y_j)_v$ are the j-th landmark coordinates of v and $\bar{J}_*(t)$ represents either J_a, J_b, J_c, J_d or J at time t, where missing coordinates are excluded. Therefore, given a library of prototypes V_l for action l, the *embedding sequence* is defined as follows:

$$D_{V_l,J_*}(t) = \min_{v \in V_l} d_{J_*(t)}(p_i(t), v), \tag{4.3.9}$$

where it is clearly shown the time dependentancy of D_{V_l,J_*}.

Given a set of actions $\mathcal{L}$ and the set of landmarks in (4.3.7), the meaningful sequences that can be extracted from $p_i(t)$ are defined as follows:

$$\begin{aligned} Seq_i(V_l) = \{&D_{V_l,J}(t), D_{V_l,J_a}(t), \dots \\ &\dots, D_{V_l,J_b}(t), D_{V_l,J_c}(t), D_{V_l,J_d}(t)\} \quad \forall l \in \mathcal{L}, \end{aligned} \tag{4.3.10}$$

Similarly, sequences for temporal information can be embedded as follows:

$$\begin{aligned} Seq_i(S_l) = \{&D_{S_l,J}(t), D_{S_l,J_a}(t), \dots \\ &\dots, D_{S_l,J_b}(t), D_{S_l,J_c}(t), D_{S_l,J_d}(t)\} \quad \forall l \in \mathcal{L} \end{aligned} \tag{4.3.11}$$

where the term V_l in (4.3.10) is replaced by S_l. This leads to two sets of sequences, $Seq_i(V_l)$ and $Seq_i(S_l)$, for all $l \in \mathcal{L}$.

It is worth mentioning that (4.3.8) allows the embedding to run and provide numerical results, even in presence of persistent missing data, due to the presence of $\bar{J}_*$. In other words, when missing data occurs, the sequences provided by (4.3.10) and (4.3.11) are lost only when *all* landmarks in the selected landmarks set are missing. In all other cases, i.e. at least one landmark is available for the selected landmarks set, the distance computation is successfully performed and the corresponding sequence is obtained.

In Section 4.4.6, it is proven that this approach not only preserves the classification integrity in case of occlusions, but also improves baseline performance.

Landmark Borrowing

In this section, a strategy to improve low-level features in case of persistent occlusions is provided. Equation (4.3.8) is based on $\bar{J}_*$, which only contains non missing landmarks. Thus, the resulting sequences in equation (4.3.10) and (4.3.11) are well-defined even in presence of missing landmarks. However, low-level sequences $p_i(t)$ and $p'(t)$ might still show missing values when an occlusion occur. To solve this problem, it is proposed to further exploit equations (4.3.8) and (4.3.9) to fill the missing values in $p_i(t)$ by using the knowledge contained in the pose libraries. Therefore, for a given time step t,

$$v^{\dagger}(t) = \underset{v \in V_l, l \in \mathcal{L}}{\arg\min}\, d_J(p_i(t), v), \tag{4.3.12}$$

is the *closest prototype* to the pose $p_i(t)$. Thus, the coordinates of the missing landmarks in $p_i(t)$ can be borrowed from $v^{\dagger}(t)$. Subsequently, $p'(t)$ is computed according to the usual definition. This strategy ensures to fill missing data in the low-level sequences. Moreover, no extra cost is required to perform equation (4.3.12), since the calculation can be embedded within

the one required by equation (4.3.9).

Short-time Interpolation

In this section, the problem of short-time occlusion is addressed. This problem occur when, within the considered sequence, landmarks coordinates are missing for only few frames. Although Kalman filter [130] can be applied to the detected landmarks, further processing is required to ensure that the Gaussian property holds for the considered data. Thus, the proposed strategy consists of interpolating available data, exploiting temporal consistency. In formulas, let $x(t)$ and $y(t)$ be the landmark coordinates with respect to time t, where some entries are occasionally missing (short-time), such that:

$$x : A \longrightarrow \mathbb{R}, \quad y : A \longrightarrow \mathbb{R}, \quad A \subset \{1, \dots, T\}, \tag{4.3.13}$$

where A represents the set of frames when the landmark is detected. Then, the missing values for $t^* \in \{1, \dots, T\} \setminus A$ are defined by the *nearest-neighbour* as follows:

$$x(t^*) = x\left(\hat{t}\right), \quad y(t^*) = y\left(\hat{t}\right), \quad \text{s.t.} \quad \hat{t} = \underset{t \in A}{\arg\min} \|t^* - t\|_2. \tag{4.3.14}$$

Given the simplicity of this solutions, it has been implemented in all proposed methods, including the baseline, as a simple and reasonable quick solution for short-time missing data.

Occlusions Augmentation

As a final strategy for occlusion-handling, synthetically occluded sequences are added in the training phase. Specifically, training sample can be persistently occluded by randomly removing some landmarks according to a binary Bernoulli distribution $\mathcal{B}(p)$ where $p = 0.5$. This strategy has been

implemented right after landmark detection. To preserve the integrity of the system, landmarks 2 and 9 have been not occluded to allow equation (4.2.4) to be well defined. This strategy aims to train resulting network with data that present random occlusions, enabling the network to learn a more general representation. It turns out that this strategy provides additional robustness to occlusions. In fact, despite the implementation of the other strategies, it is crucial to also effectively learn how the low and high-level features might change when different occlusions occur.

4.3.3 Classification Step

For fair comparisons with the proposed baseline methods, MSTLM-FCN is again used for the classification step. Depending on the input features, the classification step can focus on different motion aspects. Specifically, three sets of sequences are defined by using the proposed low-level and high-level features, as follows:

1. **Spatial-attention sequences**: these are formed by combining $p_i(t)$, $p'_i(t)$ and $Seq_i(V_l)$.

2. **Temporal-attention sequences**: these are formed by combining $p_i(t)$, $p'_i(t)$ and $Seq_i(S_l)$.

3. **Spatio-temporal-attention sequences**: these are formed by combining $p_i(t)$, $p'_i(t)$, $Seq_i(V_l)$ and $Seq_i(S_l)$.

For an overview of the ActionXPose processing, in Figure 4.4 the general pipeline of the proposed algorithm is provided.

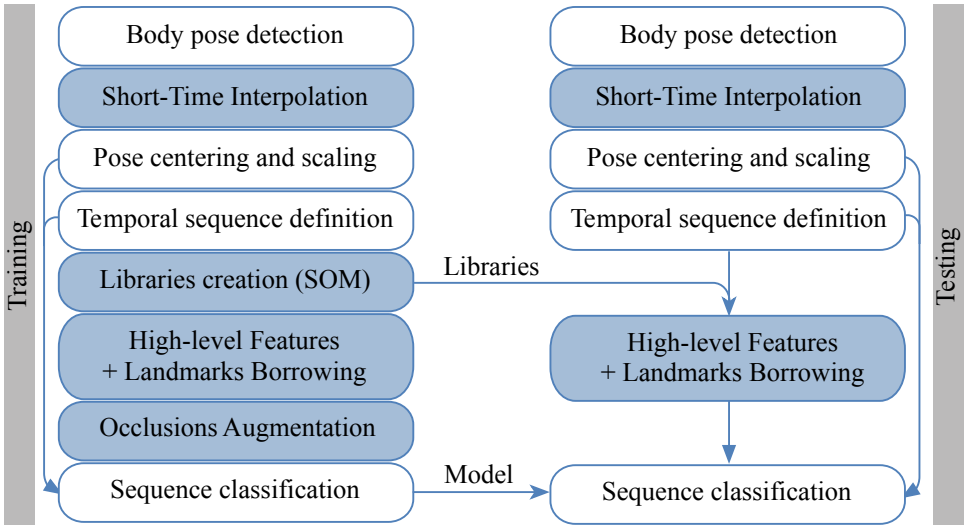


Figure 4.4: Proposed training and testing ActionXPose pipelines. The occlusions-handling steps are depicted in blue.

4.4 Experiments

4.4.1 ISLD Dataset

In this section, a new dataset for the pose-based HAR, named ISLD, is proposed. This dataset was recorded within the Intelligent Sensing Lab. Single-target CCTV-like clips, according to 18 predefined posture-related action classes, were collected. Participants were free to perform the actions according to their understanding of the class labels and no example clips were provided. Recording viewpoints were predefined, to ensure that enough viewpoints were covered. Specifically, samples were recorded from up to 5 different viewpoints, namely *front, front-left, front-right, left* and *right*. The 18 proposed actions, performed multiple times by 10 actors, were recorded with a static RGB camera. Overall, ISLD contains 907 different time windows. For each time window, only one target is visible, performing a single action. 10 examples from the ISLD dataset are shown in Figure 4.5.

ISLD samples has been pre-processed by OpenPose to extract human

Figure 4.5: Examples from the proposed ISLD dataset. Human actions of 10 subjects are recorded using a static camera, from different viewpoints.

poses. To increase the number of samples for the classification requirements, data augmentation was also performed on training data. In fact, deep learning methods usually require a great amount of training data to perform well. For example, in image recognition, *cropping* or *rotating* images are common practice to augment dataset samples in order to meet deep learning algorithms conditions [131]. In speech recognition, it is also common to add noise to training samples for the same purpose [132].

Inspired by the above-mentioned methods, the first proposed data augmentation technique is named *pose-flipping*. It consists of flipping poses along the vertical axis passing through the root landmark. This causes that the performed action looks mirrored, exploiting the left/right body symmetry. In Figure 4.6, viewpoint composition rates for the ISLD dataset are provided, showing that pose-flipping balances the left/right viewpoint rates.

The second proposed method for data augmentation is named *pose-*

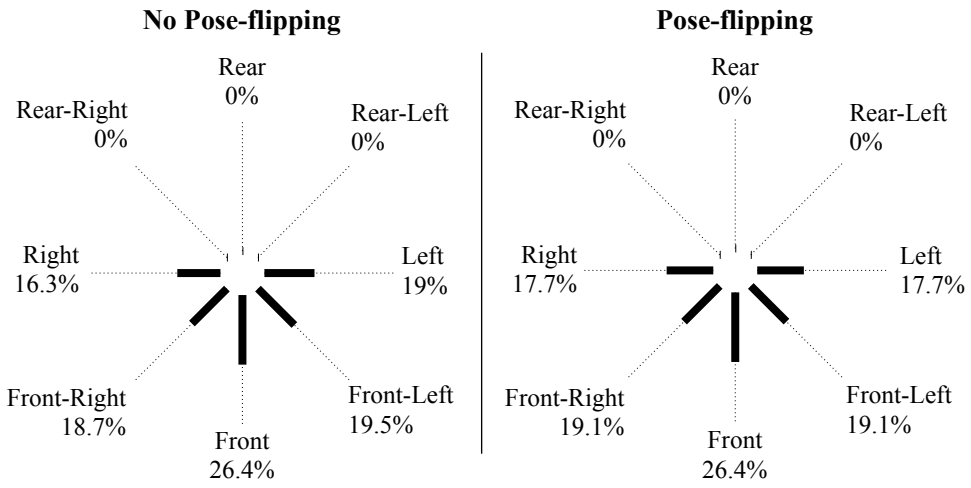


Figure 4.6: ISLD viewpoints composition rates in the cases of no pose-flipping and pose-flipping data augmentation. Pose-flipping is useful for doubling the number of samples and balancing viewpoints composition.

noising, which consists of adding Gaussian noise to the landmark coordinates, i.e. $\mathcal{N}(0, \sigma^2)$ with 0 mean and σ standard deviation. In this work, $\sigma = 0.2$ is empirically chosen for all experiments unless otherwise specified. Specifically, let z be the number of times that training data are used to create additional noisy samples. Thus, if $z = 0$, no noisy samples were created. If $z = 1$, all training samples were used *once* to create noisy samples.

In conclusion, after applying the proposed data augmentation, the ISLD dataset consists of up to 5598 samples, as shown in Table 4.1. Figure 4.6 shows that pose-flipping not only doubles the available data, but also balances left/right viewpoint rates.

4.4.2 Experimental Settings

Traditional Setting

In this setting, the training and testing phases are entirely based on the ISLD dataset. In particular, ActionXPose was trained on actors $[1, 2, 3, 4]$, validated on actors $[5, 6, 7]$ and tested on actors $[8, 9, 10]$. Regarding hyperparameters, $\sigma = 0.2$ and $z = 1$ for augmenting training data.

Table 4.1: ISLD action composition. The number of samples for each action are provided in the cases of no/yes data augmentation, by setting $z = 2$.

	Data Aug.			Data Aug.	
Label	**No**	**Yes**	**Label**	**No**	**Yes**
bend	40	240	jump	40	240
box	80	480	kick	42	252
check-watch	40	240	jump-in-place	48	288
cross-arms	40	240	point	40	240
get-up	40	240	run	40	240
hand-clap	40	240	head-scratch	40	240
one-hand-wave	40	240	sit-down	40	240
two-hands-wave	40	240	stand	350	1050
jog	36	216	walk	72	432

Multi-Datasets Augmentation Setting

In this setting, ISLD training data was considered alongside additional datasets to better leverage the deep learning generalisation ability. Since the *stand* action is already well covered by ISLD, no further data augmentation was required for this action. However, other classes are not so well represented and additional data can be helpful. Because data collection for HAR is often expensive and time-consuming, other available datasets were revised, starting from the popular UCF101 [31] and HDMB51 [117] datasets. For these two datasets, most of the video samples were collected from YouTube and movies. The camera was often too close to the target, capturing only the target's face or hands. Moreover, most samples in these datasets show low-resolution, unlabelled, multiple-target frames where the subjects perform different actions. Furthermore, most of the actions are strongly related to the context rather than to the human posture. Last but not least, as shown in Figure 4.7-(Top), if OpenPose is used to pre-process these datasets, the overall performance is too low to be a reliable source for the proposed 2D pose-based HAR. Figure 4.7-(Bottom), shows the OpenPose detection rate for UCF101 and HDMB51, supporting these conclusions.

In contrast to UCF101 and HDMB51, CCTV-like recordings often show

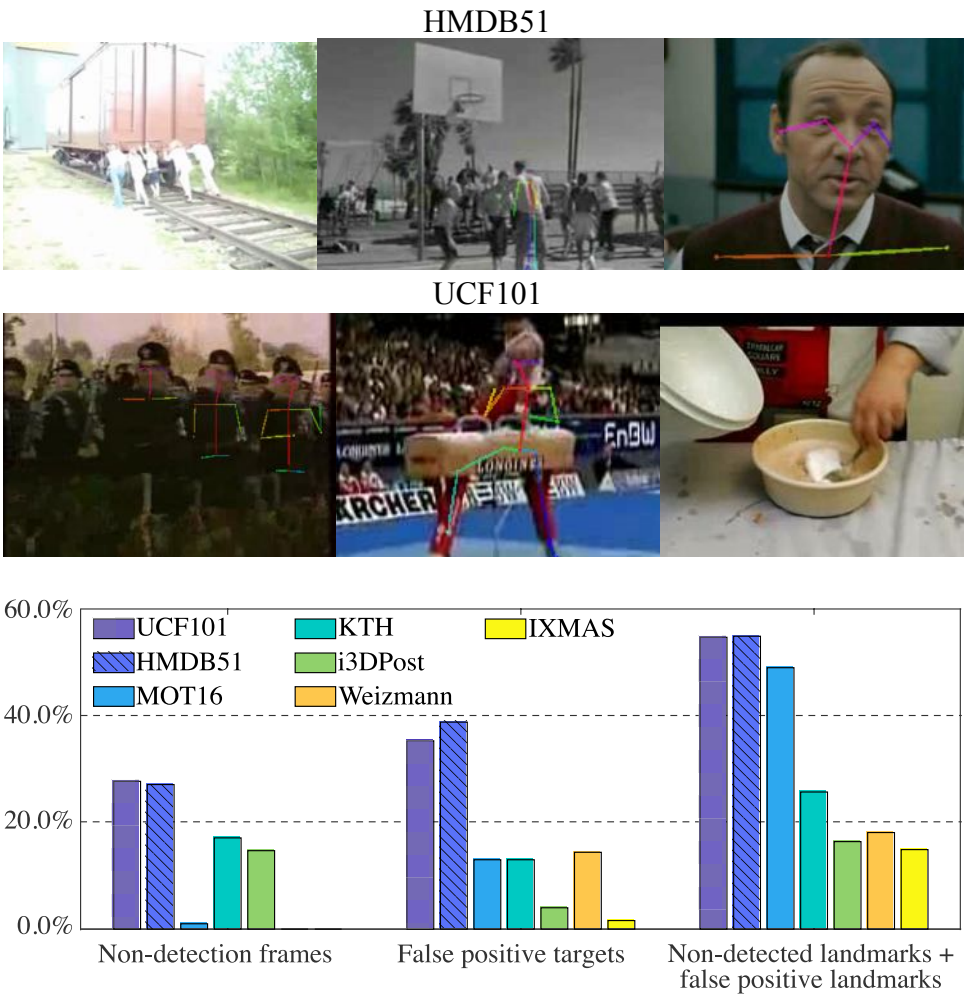


Figure 4.7: (Top) Screenshots from HMDB51 [117] and UCF101 [31]datasets processed by OpenPose. The pictures show false negatives, missing landmarks, false positives and very limited views of the target body. (Bottom) OpenPose performance on different datasets. *Non-detection frames* measures the percentage of frames in which detections do no occur. *False positive targets* measures the percentage of the root landmark confidence below the threshold 0.5. Considering those targets with the root landmark confidence higher than 0.5, *Non-detected landmarks + false positive landmarks* considers the percentage of non detected landmarks (confidence = 0), plus false positive landmarks ($0 <$ confidence < 0.5).

full-body targets, where OpenPose works well. Figure 4.7-(b) also shows the performance on a famous dataset for tracking in public environments, i.e. MOT16 [133], and other traditional datasets, i.e. KTH [7], IXMAS [9], Weizmann [6] and i3DPost [8]. It turns out that OpenPose performs considerably better on these traditional datasets. Moreover, these datasets include fully-labelled single-target clips, which simplifies the processing. In fact,

single-target data does not require tracking as pre-processing step. Motivated by these considerations, additional training data were collected from the Weizmann, i3DPost, KTH and IXMAS datasets, defining the four-in-one MPOSE dataset, by merging them all together. Moreover, the MPOSE class labels were selected to be consistent with the ISLD's labels. Overall, MPOSE contains 4160 single-target video clips, distributed over 17 action classes (the *stand* action is excluded) and performed by 53 human actors. Table 4.2 reports the MPOSE action composition for the default case (no data augmentation) and the pose-flipping/pose-nosing case. Pose-noising has the potential to indefinitely raise the total number of samples. However, for the MPOSE dataset, it turns experimentally out that $z = 2$ is a good choice for the pose-noising parameter. It is worth noticing that, when data augmentation is applied, MPOSE contains a significantly higher number of samples than UCF101 and HMDB51, which contain 13320 and 6766 samples, respectively. In terms of viewpoints, Figure 4.8 shows the viewpoint composition and the effect of pose-flipping in balancing left/right viewpoints rate.

Cross-Dataset Setting

In this setting, the MPOSE dataset was used for training and validation, while the whole ISLD dataset was used for testing. Therefore, the purpose of this test was to measure ActionXPose cross-dataset performance. In this setting, since MPOSE does not contain data for the action *stand*, the stand action is neglected from ISLD as well.

4.4.3 Implementation

Simulations were conducted on Ubuntu 16.04 running on a Dell Inspiron 15 5000 with four core Intel i7, and mounting an embedded Nvidia GeForce GTX 1050. Hyperparameters, such as number of epochs and batch size,

Table 4.2: MPOSE action composition. The number of samples for each action are provided in the Default case (no data augmentation) and after applying pose-flipping and pose-noising.

	Data Aug.			Data Aug.	
Label	**No**	**Yes**	**Label**	**No**	**Yes**
bend	193	1158	jump	73	438
box	517	3102	kick	120	720
check-watch	120	720	jump-in-place	73	438
cross-arms	120	720	point	120	720
get-up	120	720	run	474	2844
hand-clap	396	2376	head-scratch	120	720
one-hand-wave	193	1158	sit-down	120	720
two-hands-wave	407	2442	stand	0	0
jog	400	2400	walk	594	3564

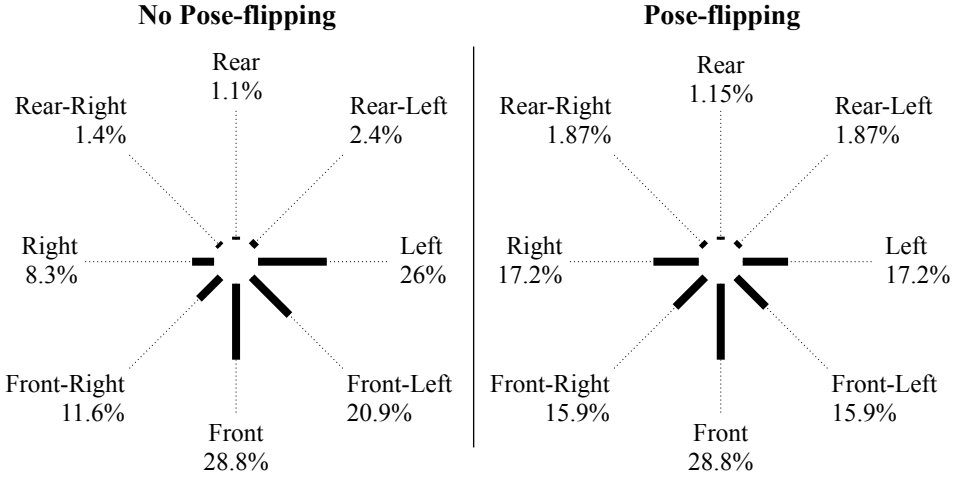


Figure 4.8: MPOSE viewpoint composition in both cases of no pose-flipping and pose-flipping.

were chosen by applying the early-stopping method to the validation sets. For the detection phase, OpenPose model is based on COCO [134], and updated versions of this detector can provide up to 25 body landmarks, 70 face landmarks, 42 hands landmarks and 6 feet landmarks for each target. However, only 14 body landmarks were exploited. Specifically, 5 out of 25 body landmarks represent *nose*, *left eye*, *right eye* and *left* and *right ears*. Therefore, these additional landmarks were averaged by defining a *head* landmark. Thus, the set of considered body landmarks is $J = \{1, \ldots, 14\}$, as described

in Figure 1.6. Finally, regarding ActionXPose coding, feature computations were conducted in MATLAB, while the classification was performed using the Keras implementation of MLSTM-FCN, provided by [129].

4.4.4 Results

In this section, ActionXPose performance on the three set of features defined in Section 4.3.3, i.e. Spatial-attention, Temporal-attention and Spatio-temporal-attention are provided. Simulations were conducted for the three experimental settings defined in Section 4.4.2. Obtained results are provided in Table 4.3 and compared with the baselines and the state-of-the-art. Regarding the Traditional experimental setting, the action classes are unbalanced due to the presence of the *stand* action. Thus, results for this setting were normalised using the total number of clips per action.

Overall, ActionXPose features outperform the baselines in almost all tests. Moreover, in the Multi-Datasets Augmentation experimental setting, additional training data improves the results obtained in the Traditional experimental setting. This is mainly due to the higher generalization degree obtained by providing additional MPOSE data during the training phase. In the Cross-Dataset setting, obtained results shows that MPOSE does not contain enough data variability to fully meet the ISLD requirements. However, it surprisingly covers most of the actions, confirming to be a good pre-training source of data. In Figure 4.9, the confusion matrix, as defined in Section 2.4, obtained in the Cross-Dataset setting is provided.

4.4.5 Ablation Study

In this section, different combinations of low and high-level features are considered. This ablation study was conducted on the MPOSE dataset. The goal of this study is to assess the contribution of each considered set of features. The cross-validation setting was used as standard method to

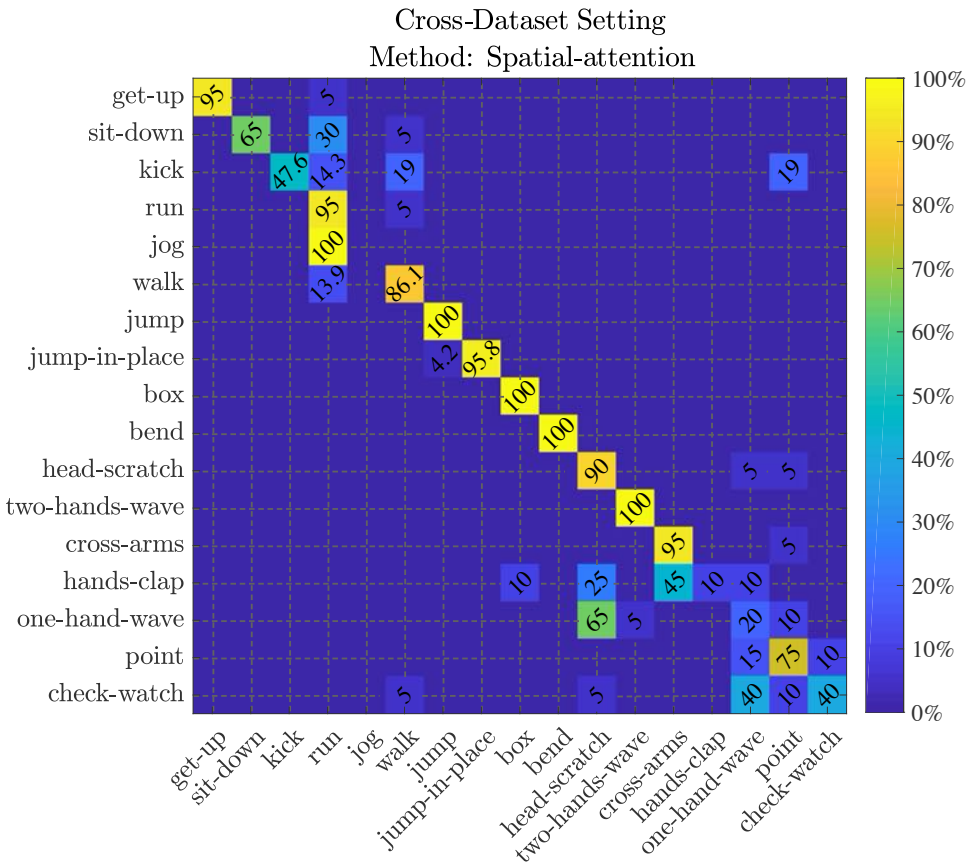


Figure 4.9: Confusion matrix obtained for the Spatial-attention method in the Cross-Dataset Setting.

Table 4.3: Accuracy results (%) for three experimental settings, i.e. Traditional (Trad.), Multi-Datasets Augmentation and Cross-Dataset. $|\mathcal{L}|$ represents the number of considered actions.

Settings	**Traditional**
$\|\mathcal{L}\|$	**18 / 17**
Baseline A [69] + [129]	92.44 / 88.88
Baseline B [68] + [129]	81.77 / 73.70
Baseline C [68] + [69] + [129]	91.99 / 87.96
Spatial-attention	91.55 / 86.24
Temporal-attention	94.22 / 91.74
Spatio-temporal attention	95.11 / 92.73
Settings	**Multi-Datasets Augmentation**
$\|\mathcal{L}\|$	**18 / 17**
Baseline A [69] + [129]	93.77 / 91.07
Baseline B [68] + [129]	79.55 / 75.00
Baseline C [68] + [69] + [129]	84.00 / 80.73
Spatial-attention	96.00 / 96.33
Temporal-attention	92.88 / 89.29
Spatio-temporal attention	96.44 / 95.58
Settings	**Cross-Dataset**
$\|\mathcal{L}\|$	**17**
Baseline A [69] + [129]	93.77
Baseline B [68] + [129]	79.55
Baseline C [68] + [69] + [129]	84.00
Spatial-attention	96.00
Temporal-attention	92.88
Spatio-temporal attention	96.44

measure accuracy performance [57, 58, 125, 126, 135–138]. Specifically, ActionXPose methods were tested by using an action-based cross-validation setting with 10 foldings. This approach stabilises the number of samples per action across different foldings, in case of imbalanced classes. Pose-flipping and pose-noising were applied to training samples, while validation and testing samples were not augmented.

The following ablated methods were considered:

$$
\begin{array}{ll}
\textbf{A} & p(t) \\
\textbf{B} & p'(t) \\
\textbf{C} & D_{V_l,J}(t),\ \forall l \in \mathcal{L} \\
\textbf{D} & [D_{V_l,J_a}(t),\ \ldots,\ D_{V_l,J_d}(t)],\ \forall l \in \mathcal{L} \\
\textbf{ABCD} & [p(t),\ p'(t),\ Seq(V_l)],\ \forall l \in \mathcal{L} \\
\textbf{F} & D_{S_l,J}(t),\ \forall l \in \mathcal{L} \\
\textbf{G} & [D_{S_l,J_a}(t),\ \ldots,\ D_{S_l,J_d}(t)],\ \forall l \in \mathcal{L} \\
\textbf{ABFG} & [p(t),\ p'(t),\ Seq(S_l)],\ \forall l \in \mathcal{L} \\
\textbf{all} & [p(t),\ p'(t),\ Seq(V_l),\ Seq(S_l)],\ \forall l \in \mathcal{L}
\end{array}
\tag{4.4.1}
$$

where the i-subscript has been omitted for convenience. It is remarked that **ABCD** and **ABFG** correspond to the Spatial-attention and Temporal-attention methods, respectively, defined in Section 4.3.3. Similarly, **all** corresponds to the Spatio-temporal-attention method. The new nomenclature in (4.4.1) better highlights the method compositions in terms of features.

Cross-validated results for these methods are shown in Table 4.4, including standard deviations, for the Default and $\sigma = 0.2, z = 2$ cases. It turns out that, in the latter case, the performance is superior due to the data augmentation provided by pose-flipping and pose-noising.

In order to measure the significance of the obtained performance, paired t-tests with $\alpha = 0.05$ were conducted for the $\sigma = 0.2, z = 2$ case. Since it is expected that the more features involved in the learning process, the higher the averaged accuracy, a one-tail paired t-test was chosen whenever possible, e.g. the two methods do not correspond to an equal number of features. In all other cases, i.e. when the number of features were the same

for both methods, a two-tail paired t-test was chosen. We report p-values for all paired methods in Table 4.4.

Overall, this ablation study revealed that the ActionXPose high-level features always bring benefits to the learning process when considered alongside low-level features, i.e. $p(t)$ and $p'(t)$. Moreover, the **all** method is not necessarily the best. This might suggests that, when too many features are involved, it is difficult for the network to effectively extract useful information. Indeed, **ABCD** and **ABFG** are significantly the best methods. This could be possibly due to the *curse of dimensionality* which occur when the number of features is too high with respect the number of available training samples [13]. This study highlighted the importance of data augmentation, which considerably improves the *Default* case performance. On the other hand, it turns out that it is advantageous to find a trade-off between number of features and accuracy, to avoid the curse of dimensionality.

Overall, as expected, this study showed the importance of the low-level features for carrying most of the action knowledge. In fact, methods **A** and **B** achieve very good performance. On the other hand, as discussed in Section 4.4.6, these methods are less robust than high-level features based methods with respect to occlusions. Therefore, high-level based methods provide the required robustness, along with clear advantages in terms of accuracy.

4.4.6 Occlusions Study

In this section, the robustness of ActionXPose features to occlusions and missing data is evaluated. In particular, the contributions of the strategies proposed in Section 4.3.2 for occlusion-handling are highlighted.

The short-time occlusions strategy in Section 4.3.2 is always applied. Moreover, in all experiments, pose-flipping and pose-noising were always applied to training data, with $\sigma = 0.2$ and $z = 1$.

This study has been conducted on MPOSE dataset by using the cross-

Table 4.4: ActionXPose ablation study based on MPOSE dataset. Performance of different methods is provided on average (AVG) for 10 cross-validation foldings, reporting obtained standard deviations (STD). The **Features** column reports the actual number of features available for each method. Regarding hyper-parameters, **Default** and $\sigma = 2, z = 2$ cases are reported.

		Default		$\sigma = 0.2, z = 2$	
Method	**Features**	**AVG (%)**	**STD**	**AVG (%)**	**STD**
A	28	91.47	1.68	93.87	1.26
B	28	91.72	1.20	92.64	1.43
C	17	88.90	1.50	91.88	1.68
D	68	91.31	2.03	92.11	1.44
ABCD	141	92.95	0.89	95.48	1.34
F	17	77.53	3.29	78.47	5.77
G	68	84.28	2.36	86.60	2.05
ABFG	141	93.69	1.61	95.43	1.14
all	226	92.54	1.87	94.44	0.81

Table 4.5: p-values (significance) obtained by conducting a paired T-test on each pair of methods considered in the ablation study, in the case of $\sigma = 0.2, z = 2$. If the p-value $> .05$, the two method results are not significantly different to each other.

p-values (significance) for $\sigma = 0.2, z = 2$								
	B	**C**	**D**	**ABCD**	**F**	**G**	**ABFG**	**all**
A	.025	.009	.000	.000	.000	.000	.003	.009
B	n/a	.139	.211	.000	.000	.000	.000	.001
C	n/a	n/a	.690	.000	.000	.000	.000	.000
D	n/a	n/a	n/a	.000	.000	.000	.000	.000
ABCD	n/a	n/a	n/a	n/a	.000	.000	.919	.030
F	n/a	n/a	n/a	n/a	n/a	.001	.000	.000
G	n/a	n/a	n/a	n/a	n/a	n/a	.000	.000
ABFG	n/a	n/a	n/a	n/a	n/a	n/a	n/a	.001

validation approach already discussed in the previous section. Since MPOSE data contains only self-occlusions, more challenging occlusions were simulated by explicitly removing landmarks from the testing data. This strategy is fast and effective, it does not require any time-consuming video editing, and provides similar results as assumed the occlusions are in the video data. Inspired by landmark subsets in (4.3.7), 6 different groups of landmarks were

purposely neglected, i.e.

$$
\begin{aligned}
J_a^* &= \{4, 5\} && \text{(Right Arm)} \\
J_b^* &= \{7, 8\} && \text{(Left Arm)} \\
J_c^* &= \{10, 11\} && \text{(Right Leg)} \\
J_d^* &= \{13, 14\} && \text{(Left Leg)} \\
J_{a,b}^* &= \{4, 5, 7, 8\} && \text{(Both Arms)} \\
J_{c,d}^* &= \{10, 11, 13, 14\} && \text{(Both Legs)}
\end{aligned} \tag{4.4.2}
$$

It is worth emphasising that the baseline methods are strongly numerically affected by the proposed occlusions. In other words, such occlusions create persistent missing data, and thus persistent missing features. On the other hand, the ActionXPose high-level features are more numerically robust, due to the definition of the embedding distance in (4.3.8). Specifically, when such occlusions occur, the proposed high-level features only slightly change their values, rather than being completely lost.

The first experiment (Figure 4.10-Top) consists of occluding testing data, without performing neither occlusions augmentation nor landmarks borrowing techniques. Thus, the trained networks were not prepared to face such occlusions. As expected, the baseline methods are strongly less robust than ActionXPose features. In contrast, all methods that include high-level features achieve much better performance due to the robustness provided by equations (4.3.8) and (4.3.9). In particular, the proposed Spatio-temporal attention method remarkably outperforms the baselines in all the occlusion cases.

In the second experiment (Figure 4.10-Middle), occlusions augmentation was enabled. In this case, since the training data include synthetically occluded data, the resulting networks are much more robust to occlusions. In this case, baseline methods are also expected to be more robust since the

trained network is prepared to deal with the missing features carried by low-level sequences. However, again, high-level features outperform the baselines in all cases.

In the third and last experiment (Figure 4.10-Bottom), both occlusions augmentation and landmarks borrowing were enabled. While the occlusions augmentation include synthetically occluded data into training data, the borrowing landmark technique is able to fill the gaps due to occlusions in the baseline features. To perform this experiment, training and validation data were firstly occluded by the occlusions augmentation technique, while testing data were occluded with the proposed equation (4.4.2). Then, all low-level and high-level features for training, validation and testing data were computed considering the borrowing landmarks technique. The first effect of this processing is that performance and robustness globally further increase. However, again, the proposed ActionXPose features outperform the baselines in all occlusion cases.

In conclusion, the results of this study showed that the proposed occlusion-handling techniques are advantageous and provide complementary improvements over the baselines in terms of robustness to occlusions.

4.4.7 Performance on Traditional Datasets

In this section, ActionXPose results on the KTH and i3DPost datasets are provided. These tests were conducted to allow comparisons between the proposed method and other state-of-the-art methods. Since KTH and i3DPost include specific challenges, such as multiple viewpoints, zooming in/out, moving cameras, and variable target-camera proximity, this test can also show ActionXPose robustness against these challenging conditions.

The tests on the KTH dataset were performed under two experimental settings. The first is the *Split* setting, where training, validation and testing samples are predefined by the original author in [7]. The second is the

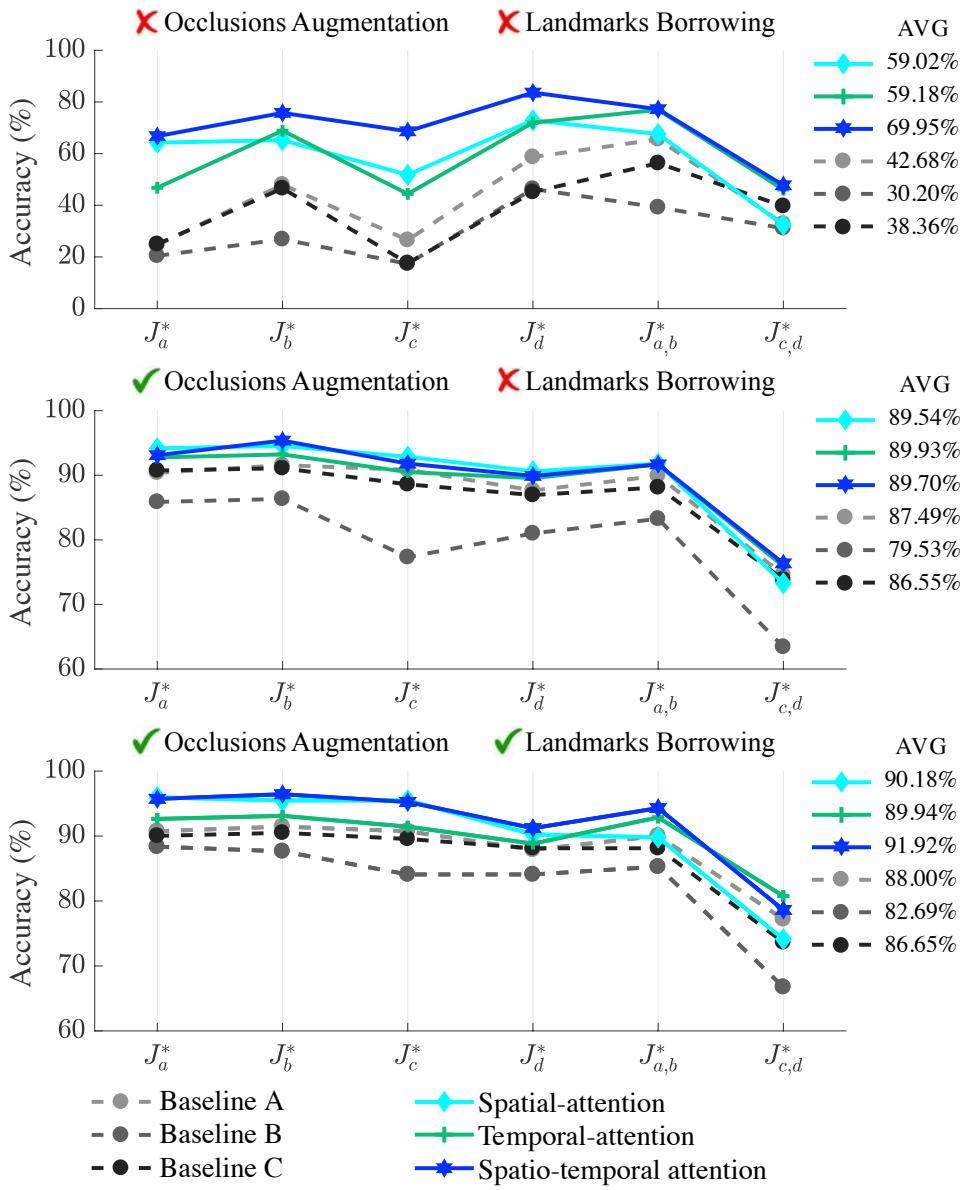


Figure 4.10: Occlusion Study results, provided on average over 10 cross-validation foldings. AVG represents the averaged results over the six occlusion cases. (Top) Performance with no data-augmentation nor borrowing landmarks. (Middle) Performance with data-augmentation but no borrowing landmarks. (Bottom) Performance with data-augmentation and borrowing landmarks.

Leave-One-Actor-Out (LOAO) setting, where multiple tests are conducted by using each actor as testing actor and averaging obtained results. Table 4.6 shows the results for both these experimental settings. Data augmentation parameters were empirically chosen and fixed for all tests as $z = 0$ and $\sigma = 0.2$.

Table 4.6: Accuracy results (%) for KTH and i3DPost. For KTH dataset, two settings are reported, namely Split and LOAO. For i3DPost dataset, LOAO results are obtained for $|\mathcal{W}| = 8$, where $|\mathcal{W}|$ represents the number of considered viewpoints. The ($\circ$) denotes the case $|\mathcal{W}| = 2$.

	KTH		**i3DPost**
Method	**Split**	**LOAO**	**LOAO**
Spatial-attention	90.50	**99.04**	98.95
Temporal-attention	90.15	98.03	98.95
Spatio-temporal attention	89.80	98.26	99.47
Baseline A [69] + [129]	88.06	98.91	97.39
Baseline B [68] + [129]	83.19	96.29	95.30
Baseline C [69] + [68] + [129]	86.44	96.67	99.47
Kovashka et al. [75]	**94.50**	n/a	n/a
Zhang et al. [76]	94.10	n/a	n/a
Ji et al. [77]	90.20	n/a	n/a
Almeida et al. [78]	n/a	98.00	n/a
Vrigkas et al. [79]	n/a	98.30	n/a
Liu et al. [80]	n/a	93.80	n/a
Raptis and Soatto [81]	n/a	94.50	n/a
Jiang et al. [82]	n/a	95.77	n/a
Gilbert et al. [83]	n/a	95.70	n/a
Chapter 3, Framework B	n/a	n/a	**99.60**
Castro et al. [126]	n/a	n/a	99.00($\circ$)
Iosifidis et al. [57]	n/a	n/a	98.16
Azary et al. [58]	n/a	n/a	92.97
Hilsenbeck et al. [125]	n/a	n/a	92.42

The ActionXPose high-level features outperform the baseline methods in all settings. Moreover, in the Split setting, ActionXPose performance is among the state-of-the-art. In the case of the LOAO setting, ActionXPose outperforms other state-of-the-art methods.

Regarding i3DPost, this dataset is usually tested under the LOAO setting. i3DPost is specifically designed to perform multi-viewpoints HAR. In fact, it includes video clips recorded from 8 different viewpoints. The results are given in the challenging multi-viewpoints case, i.e. training and testing data include all viewpoints data. ActionXPose results are summarised in Table 4.6 and compared with the state-of-the-art. In this test, ActionXPose outperforms the baseline methods, achieving performance which are among the state-of-the-art.

In conclusion, these tests are particularly suitable for highlighting the effectiveness of pose-based HAR in comparison with the traditional methods. In fact, such excellent results were obtained by using 2D human poses only, while the other state-of-the-art methods exploited RGB data or other sophisticated data sources such as human silhouette. In particular, it turns out that 2D pose-based HAR can achieve similar, and sometimes superior, performance than traditional RGB based methods.

4.4.8 Computational Effort and Execution Time

In this section, the computational speed evaluation is provided for each of the most important step required by ActionXPose.

The first, computationally expensive step is due to the pose detector. The performed simulations showed that the body pose detector is the bottleneck for the entire processing. However, it is claimed to be a real-time detector [12] when hardware requirements are properly satisfied. This statement is supported by the performed experiments.

Moreover, let separately consider the *training* and *testing* phases.

Training phase

In the training phase, the most intense step (excluding the body pose detection) is the library creation step, where the SOM clustering method is performed. Following the same notation as in Section 4.3.1, the SOM computational complexity can be estimated as $\mathcal{O}(N^2m^2)$ [139], where N is the number of considered samples and m is the input dimensionality. This argument shows the importance of considering the PCA as a dimensionality reduction technique before running the SOM, in reducing the m value.
Another important step is sequence embedding step, which has a complexity of $\mathcal{O}(n)$, where n is the number of prototypes in the considered library (see Section 3.4.4). Thus, for the considered $n = 64$, as discussed in Section

4.3.1, the embedding process is remarkably fast, i.e. $\approx 9 \times 10^{-5}$ SPF, which corresponds to around 10^4 FPS. Even considering that ActionXPose needs to run two embedding processes (for the spatial and temporal libraries), the embedding process is still very fast.

Testing phase

The testing phase is less complex and faster than the training phase. In this phase, the most complex step (excluding the body pose detection) is the embedding step for computing high-level features and performing the landmark borrowing. As already discussed in the training phase, the complexity of this operation is $\mathcal{O}(n)$, where n is the number of prototypes in the considered library. In Figure 4.11, the testing phase pipeline is depicted, reporting the obtained performance in terms of SPF and FPS for each required step. These results have been obtained by ActionXPose **all** method. It can be seen that, excluding the body pose detection step (performed by OpenPose), the proposed processing is remarkably fast. In fact, the processing time magnitude varies between 10^{-5} and 10^{-4} SPF. However, OpenPose is the bottleneck of the entire processing, running at a time magnitude of 10^{-2} SPF. Overall, ActionXPose can produce an output in around 3.2×10^{-2} SPF, which corresponds to 31.23 FPS.

4.4.9 Varying Video Quality Study

In this section, additional insights about the proposed method robustness to different frame resolutions, colour channels, number of frames per second, mega-bits-per-second (mbits/s), actual body size and frame quality rate. In fact, it is reasonable to expect that different video qualities, in terms of the above-mentioned indicators, might result in different OpenPose performance, which in turn can affect ActionXPose performance.

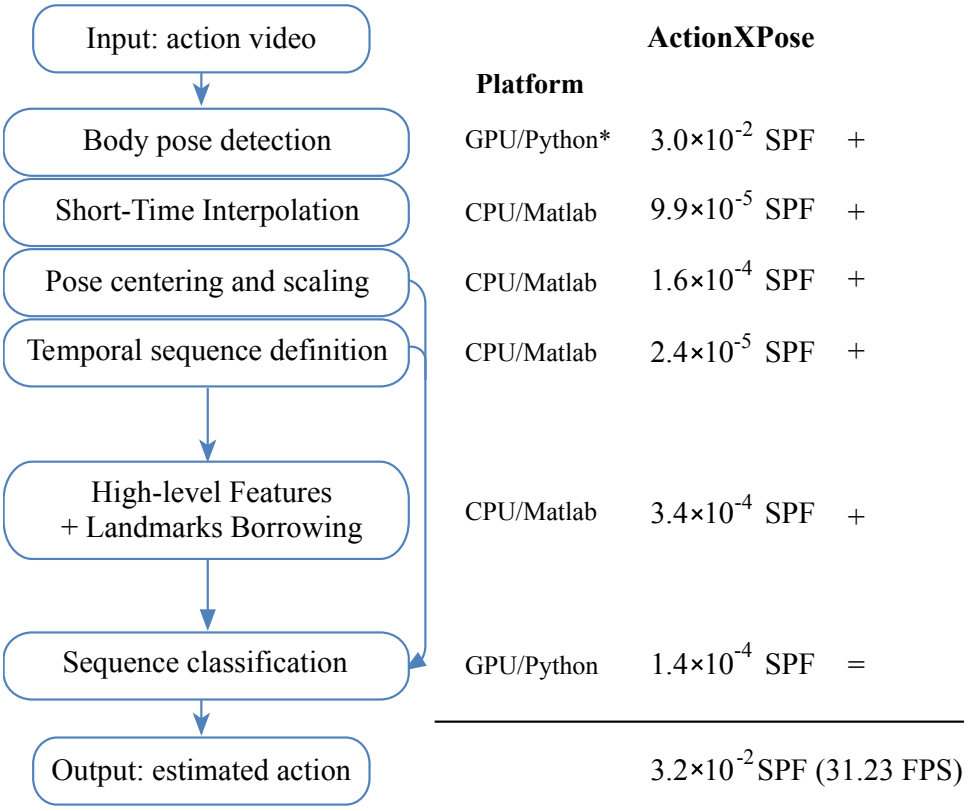


Figure 4.11: Execution time for each ActionXPose step. The execution time for each processing step is independently measured in Seconds-Per-Frame (SPF). The cumulative SPF is computed and transformed in Frame-Per-Second (FPS). The CPU/Matlab platform is a Windows 7 workstation, 64-bit, with a CPU Intel Core i5-6600 @ 3.30GHz 3.30GHz, 16 GB of RAM, running Matlab implementations. The GPU/Python platform is a Linux Ubuntu 16.04 workstation, 64-bit, with a CPU processor Intel i7 @ 3.1GHz 3.1GHz, 8 GB of RAM, hosting a NVIDIA GEFORCE GTX 1050 graphics card with 4GB GDDR5, running Python implementations. GPU/Python* represent the workstation suggested by [12] to achieve 32 FPS.

As first insight, in Table 4.7, the datasets used in this chapter are compared in terms of these common video indicators. The variety of conditions shown in Table 4.7, compared with the performance presented in previous sections, demonstrates that ActionXPose performance is stable across different conditions.

As an additional study, further experiments on ISLD were conducted, under the Dataset Augmentation Setting presented in Section 4.4.2.

The first goal was to assess the impact of varying frame sizes and body sizes on ActionXPose performance. To this purpose, the original ISLD frame

Table 4.7: Video quality comparison, reporting for each used dataset, colour channels (Chan.), frame-per-second (FPS), mega-bits-per-second (mbits/s), Frame Size and averaged Body Size.

Dataset	Chan.	FPS	mbits/s	Frame Size	Body Size
Weizmann	RGB	25	15.55	180×144	65×93
i3DPost	RGB	25	5.18	960×540	384×408
IXMAS	RGB	19	1.9	390×291	136×73
KTH	mono	25	0.89	160×120	82×106
ISLD	RGB	25	47.97	1920×1080	403×557

size was repeatedly reduced by a factor of 5. Each time, the resulting clips were saved in AVI format with the Motion JPEG 2000 encoder provided by MATLAB, with a quality threshold of 95%. Obtained results are reported in Figure 4.12-Top.

The second goal was to assess the impact of varying Motion JPEG quality rates on ActionXPose performance. Therefore, the frame size was set to a reasonable value, i.e. 192x108 pixels, and then repeatedly reduced the quality threshold from 95% to 35%. Obtained results are reported in Figure 4.12-Bottom.

Overall, conducted tests showed that the body size reduction is slightly related to a reduction of OpenPose performance. Similarly, ActionXPose performance slightly reduces. However, the loss in performance is limited (around 3%). Similarly, the frame quality rate slightly worsens OpenPose performance. However, this further false negative increment seems to have a limited impact on ActionXPose performance.

In conclusion, these results suggest that ActionXPose is robust to the studied working conditions changes.

4.4.10 Critical Analysis

In this section, the results presented in this chapter are critically analysed and compared to the aims and objectives set in Section 1.3.

Objective 4 and 5 have been successfully achieved. In fact, the method

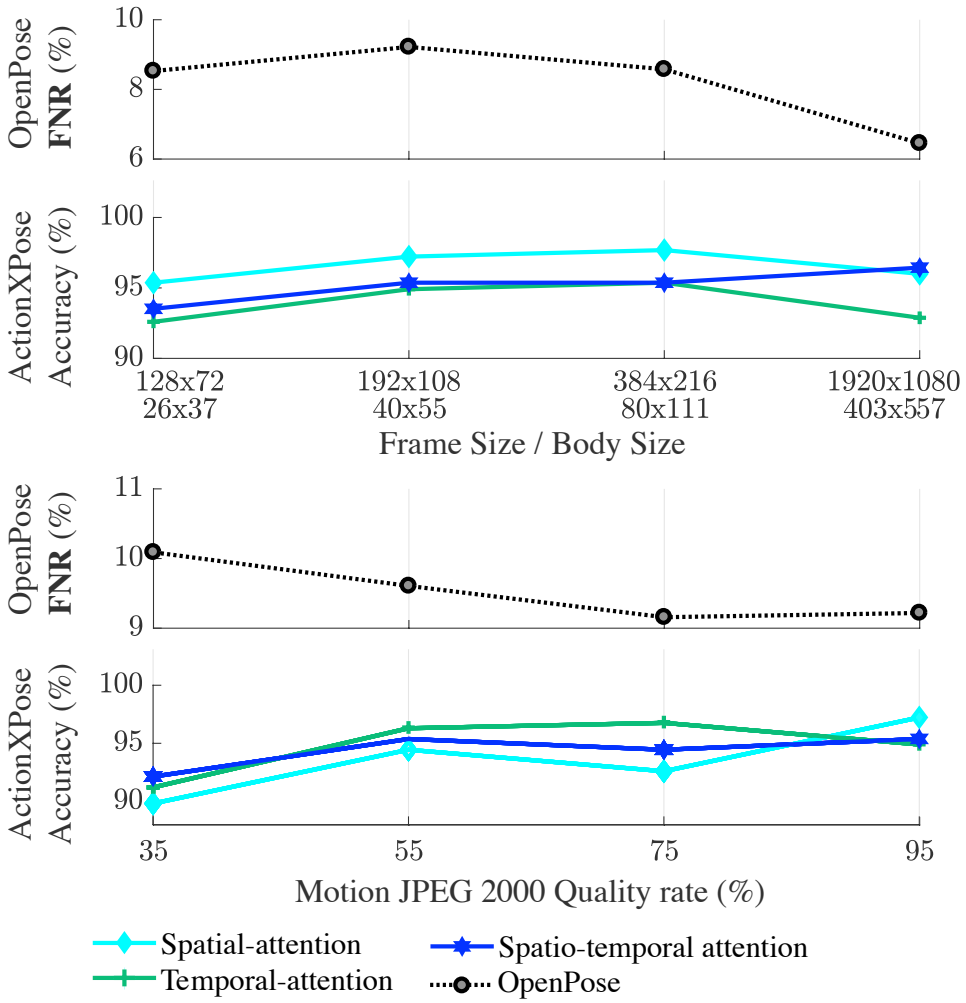


Figure 4.12: Frame size, body size and quality rate impact on OpenPose in terms of False Negative Rate (**FNR**), compared with ActionXPose performance in terms of accuracy. (Top) Impact of the frame size and body size changes on OpenPose and ActionXPose methods. (Bottom) Impact of the frame quality changes on OpenPose and ActionXPose methods.

proposed in this chapter is based on high-level features which are computed following the embedding approach already exploited in Chapter 3. Therefore, the effective approach explored in Chapter 3 has been transformed into a solution to effectively extract features from human poses. Moreover, human poses showed the following advantages compared to human silhouettes: 1) Despite different video datasets visually show different light and background conditions, OpenPose can effectively extract homogenous pose data, allowing multi-datasets and cross-datasets experiments; moreover, poses are more

reliably extracted from different data than silhouettes; 2) human poses are much faster to be processed, once they are detected, compared to human-silhouettes; 3) human poses can be also effectively exploited for tracking, allowing the proposed method to generalise to multiple-targets. Thus, the limitations reported in Chapter 3 can be effectively solved by using human-poses.

Objective 6 has also been effectively achieved. The main limitation of using the human poses is the relatively high false-negative rate, due to adverse video conditions that, occasionally or persistently, might compromise the landmark detection. Moreover, human-poses are naturally prone to self occlusions, similarly to human-silhouettes, due to its intrinsic 2D nature. In other words, both human-poses and human-silhouettes are strongly viewpoint-dependent. Therefore, it was required to propose effective solutions to the high false-negative rate and self-occlusion issues. The occlusion-handling strategies discussed in this chapter effectively address these problems. In particular, as shown in Figure 4.10, the baseline methods are not suitable to face body limbs occlusions. In contrast, the proposed method shows greater robustness when body limbs occlusions or false detection occur.

Despite the advantages mentioned above, the experiments presented in this chapter also showed OpenPose limitations. In fact, in the case of UCF101 and HMDB51 datasets, OpenPose's low performance was one of the causes that compromised the proposed processing. As shown in Table 4.7 and Figure 4.12, frame resolution and frame quality are not a major problem themselves. In contrast, as the UCF101 and HMDB51 results suggest, the major issues were due to strong ambiguity between the target and the background. Moreover, the very small ratio between the target size and the frame resolution also compromised the detection.

The above mentioned OpenPose limitations have an impact on Action-XPose. In fact, when the link vectors $v_{2,9}$ and $v_{2,12}$ are both persistently

missing, the strategy in Section 4.2.1 is no longer well-posed. However, this case occurs only when the body trunk is persistently occluded or undetectable. In this case, even human eyes might fail in classifying posture-related actions such as those studied in this work. Such disruptive cases require ad-hoc studies which are beyond the scope of this thesis.

Another problem occurs when more semantically challenging action classes are considered for HAR; for example, those provided by UCF101 and HMDB51 datasets. Since the UCF101 and HMDB51 are *not* posture-related datasets, colour contextual information is crucial, and the poses are not informative enough to fully describe the performed action. However, non-posture-related HAR is beyond the purpose of this thesis.

Finally, two questions regarding the usage of human-poses are necessarily raised.

The first question is related to the computational effort required to preprocess data to extract posture-related information. In the case of silhouette-based HAR, in Chapter 3, the background subtraction algorithm ViBE was exploited. In this chapter, the proposed HAR method uses OpenPose to detect human poses. OpenPose requires powerful GPUs to run within a real-time performance. On the contrary, ViBE can achieve the same performance by using a simple CPU. On the other hand, as already discussed in Section 3.5, ViBE does not perform any semantic detection, while OpenPose performs human detection (classification), body limbs detection (classification), and allows multiple-human tracking. Therefore, the more powerful hardware required by OpenPose is justified by the more complete and informing outputs. For these reasons, in this thesis, OpenPose is considered more advantageous than ViBE.

The second question is related to privacy issues. On one hand, human poses-based processing increases the privacy of the monitored subjects, since RGB data can be promptly discarded right after the detection and

not memorised at any stage. On the other hand, human poses still may contain privacy-related information that can be indirectly used to identify human subjects. Nevertheless, in case of less restrictive approaches regarding privacy protection, further studies can be conducted to consider *colours* alongside *poses*, to jointly extract useful information in a multimodal system. These studies are conducted in the next chapter.

4.5 Chapter Summary

In this chapter, the proposed ActionXPose algorithm for 2D pose-based HAR has been presented, which achieves state-of-the-art performance and outperform the baselines. Proposed high-level features improve accuracy and robustness to occlusions and missing data in comparison with the baselines, which in turn are based on low-level features. In addition, this chapter proposed a new dataset for posture-related HAR in CCTV-like environment, namely ISLD dataset. This dataset was used to extensively test several variations of the proposed method, under different experimental conditions, including the interesting Dataset Augmentation and Cross-Dataset settings.

Chapter 5

POSE-DRIVEN HUMAN ACTION RECOGNITION AND ANOMALY DETECTION

5.1 Introduction

In this Chapter, the contributions discussed in Chapter 4 are extended and exploited for non-cooperative, simultaneous HAR and HAAD. Particularly, Objectives 7-9 introduced in Section 1.3 are addressed.

Similar to Chapters 3 and 4, the focus of this chapter is on posture-level HAR. In particular, the ActionXPose algorithm, which has been contributed in the previous chapter, intentionally, neglects RGB data for pose-based HAR. Thus, in this chapter, deep-learning-based HAR methods are studied, which jointly extract hidden-features from RGB and poses data. Moreover, in this chapter, the anomaly detection problem is addressed. The focus in on the normal/abnormal body posture and object-position related events. For example, a person falling, running or fighting in a scene where these events do not normally happen, it is considered as an abnormality. Similarly, a key object unexpectedly positioned or manoeuvred is considered as an abnormal event.

Therefore, in this chapter, a novel system to simultaneously perform

HAR and HAAD on posture-level human activity is contributed. It is based on pose-driven deep learning networks that rely on human pose detections and RGB data within target bounding boxes to perform Joint RGB-poses based features extraction, classification and anomaly detection.

HAR and HAAD are, conventionally, two distinct video processing stages that require different solutions. However, in this chapter, it is demonstrated that they can be performed simultaneously as two steps of a common pipeline. In other words, the proposed system is able to simultaneously detect anomalies related to the body movements and object positions while providing human action labels, to serve as an explanation of the normal/abnormal detected action. Moreover, the resulting system is easily generalisable to multiple human targets and multiple objects. The methodology followed in this chapter is threefold:

1. The ActionXPose algorithm proposed in Chapter 4 is transformed from poses to joint RGB-poses. To this end, different deep-learning architectures are proposed, and their performance in terms of HAR are studied. It is shown that joint RGB-poses networks outperform their single modality counterparts, in terms of HAR accuracy.

2. The best joint RGB-poses architecture is further studied in the context of HAAD. Particularly, the multimodal hidden features that the network can extract are exploited for semi-supervised posture-related HAAD. The contribution of RGB and pose data is compared in terms of anomaly detection accuracy. It is shown that RGB-based and joint RGB-poses based features are suboptimal for HAAD when compared to the poses-based features. In other words, poses-based features excel for HAAD, while joint RGB-poses based features excel for HAR. Thus, an effective method to train both the HAAD and HAR architectures is proposed, which aims to maximise the efficacy of RGB-poses features

for HAR and poses-based features for HAAD.

3. The role of contextual objects in the scene is also considered as an additional level of HAAD analysis. Thus, in this chapter, a method to combine objects-positions related anomaly detection with the above-mentioned posture-related HAAD is also proposed.

Extensive simulations are provided to highlight the efficacy and robustness of the proposed joint RGB-poses based HAR. In particular, three novel models for Joint RGB-poses deep-learning networks for HAR are defined, and their performance evaluated on UCF101 and MPOSE datasets. The goal is to demonstrate that joint RGB-poses based networks are more effective for HAR than their single RGB or pose counterparts. Therefore, the best model is further exploited as a deep learning feature extraction for combined HAR-HAAD.

Extensive HAAD performance evaluation is carried out on three novel datasets recorded in the Intelligent Sensing Lab, i.e. the Body Movements based Dataset (BMbD), the Multi-target Body Movements based Dataset (M-BMbD) and the Joint Body Movements and Object Position based Dataset (JBMOPbD), specifically designed for the above-mentioned problem. In particular, BMbD proposes single-target posture-related based anomalies; M-BMbD considers multi-target posture-related based anomalies; JBMOPbD further considers single-target posture-related anomalies alongside multi-object position based anomalies. Comparative results over the state-of-the-art are also provided to highlight the efficacy of the proposed method.

As already mentioned, this chapter studies are entirely based on pose-driven networks, as human poses contain most of the useful information for posture-based HAR and HAAD. However, it is shown that RGB data-based networks are less effective for HAR compared to the joint RGB-poses

ones. However, regarding HAAD, RGB data allows the resulting system to detect, in principle, abnormal activities coming from any source, such as people, cars, explosions and traffic. Moreover, RGB based methods can work in cluttered, in-the-wild and low-resolution situations. However, in the case of posture-level HAR and HAAD, this chapter provides evidence that RGB based methods are suboptimal. Moreover, when the scene semantic is given and known, i.e. human posture-level actions in an indoor scene, RGB-based methods do not explicitly consider this additional contextual knowledge. In particular, taking into account the limited availability of published implementations of state-of-the-art anomaly detection algorithms, in this chapter, two state-of-the art methods [109, 110] for autoencoders-based anomaly detections were chosen as baseline. This particular choice was motivated by: 1) autoencoders represent a well-studied and general methods to perform video-based anomaly detection; in particular, [110] is claimed to be *domain free*, suggesting that the method was developed to be a rightful choice for video-based anomaly detection from any video source; 2) [109,110] offer a ready-to-use implementation of their methods. However, despite their promising results presented in in-the-wild anomaly detection, it is shown that these architectures are not the best choice to address the challenges considered in this chapter. In contrast, a joint RGB-poses algorithm is proposed, which simultaneously perform HAR and HAAD, outperforming the state-of-the-art in terms of HAAD performance.

This chapter focuses on *semi-supervised* HAAD [14]. In other words, during the training phase, no abnormal instances are presented to the algorithm to build up abnormal expectations. Therefore, the anomaly detection algorithm is trained on the most challenging case where only normal data is available for training.

In summary, the contributions of this chapter are mainly threefold:

- A novel HAR algorithm which exploits both human poses and RGB

data;

- A novel pose-driven, bi-level, i.e. Body-Movements (BM) and Object-Position (OP), HAAD algorithm, which is integrated with the above-mentioned HAR algorithm for simultaneous HAR and HAAD;
- Three novel datasets for challenging human posture-related and object position HAAD, i.e. BMbD, M-BMbD and JBMOPbD.

The rest of this chapter is organised as follows. Section 5.2 presents the novel RGB-poses based networks for HAR. In Section 5.3, the strategy to simultaneously perform HAR and HAAD is proposed. Simulations and results are provided in Section 5.4 and conclusions are summarised in Section 5.6.

5.2 Joint RGB-poses Networks for HAR

The goal of this section is to propose the Joint RGB-poses networks for HAR. In particular, supervised learning-based structures are proposed for end-to-end recognition tasks. Therefore, the output of these networks is an action label and confidences over a pre-defined set of labels.

5.2.1 Pose Features Extraction

The first step consists of extracting pose-related information from RGB data. To this purpose, OpenPose [12] is exploited to detect body landmarks of visible targets. This detector is able to provide up to 25 body landmarks, detection confidence for each landmark. The target bounding box can be simply retrieved by the detected landmarks set.

In principle, since multiple targets can be visible in each frame, for simplicity, a Kalman's Filter based tracking systems is used to track detected poses and RGB ROIs accordingly [87].

Therefore, the tracked poses, i.e. landmark time sequences, are treated with ActionXPose with spatio-temporal-attention, as described in Chapter 4. Therefore, the output of this step consists of features in the form of multivariate time sequences, encoding different aspect of the motion, including the mutual relation between sub-body parts and full-body similarities with the learned libraries.

5.2.2 Proposed Joint RGB-poses Models

This section aims to propose three models for Joint RGB-poses based HAR. The main idea is to jointly learn from both RGB data and pose-related features extracted in Section 5.2.1, imposing a collaborative deep learning structure.

Theoretically, let I represents the aggregate information contained within RGB data, then the pose-related information P is such that $P \subset I$ (Figure 5.1), because poses are in turn obtained from RGB data. Therefore, it is reasonable to argue that learning from RGB data (which contains the whole information I) is expected to give similar, if not superior, results compared to learning from poses. However, as shown in this chapter, this is not the case, due to the deep learning limitations, e.g. it does not allow to explicitly point the network's attention over the right image regions according to a human-like perspective. Therefore, the challenge is to learn from P by using the pose-branch and from $I \setminus P$ by using the RGB-branch, maximising performance.

To adequately investigate the joint RGB-poses learning process more independently from a given deep learning structure and obtain more general conclusions, three Joint RGB-poses models, named Model A, B and C, are proposed.

Model A, B and C share a common parallel-learning structure [19], where poses-based and RGB-based information flow in two deep learning

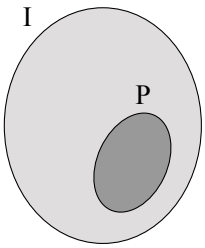


Figure 5.1: Aggregate information I provided by RGB data, including the pose-related information P.

branches. However, as opposed to the famous parallel learning approach Two-stream [140], the proposed structure is not based on two completely independent branches. Generally, independent branches are those where the final decision is made by fusing the information at the score level. However, this approach prevents from considering dependency between modalities. For example, it is reasonable to expect that poses-based learning is particularly effective for posture-related actions, while RGB data can describe finer appearance-related details, supporting the action recognition especially when poses are not enough informative. Moreover, challenging viewpoints can produce incomplete or corrupted poses that RGB data can compensate. Therefore, in these particular cases, it might be convenient to trust more RGB data than poses. In contrast, when target appearance is cluttered or unexpected, poses might be the best modality to rely on. Therefore, it is required to *dependently* discriminate data sources. Inspired by information fusion in [141], this inter-dependency is modelled by firstly process RGB data and poses separately. Subsequently, features vectors are concatenated, followed by a Fully Connected layer (FC) to learn dependency between feature vectors. Due to the back-propagation, the network can jointly optimise its weight according to the contribution of both modalities depending on the conditions.

Given the above-mentioned common learning structure, it is necessary to specify how to extract feature vectors from each modality. Regarding the

pose-branch, Models A, B and C are based on the Multivariate Long Short-Term Memory and Fully Convolutional Network (MLSTM-FCN) network [129], as discussed in Chapter 4. In contrast, regarding the RGB-branches, inspired by [19], three different networks are considered:

CRNN

RGB data is processed through four CNN and two Fully Connected (FC) layers to extract frame-level features. Subsequently, the LSTM is used to model the time dependencies of the features across frames.

ResNetCRNN

RGB data is processed by using ResNet152 [142], pre-trained on ILSVRC-2012-CLS dataset [143], to extract frame-level features. Subsequently, LSTM is used to model the time dependencies of the features across frames.

3DCNN

RGB data is processed by using two 3DCNN and 2 FC layers to both extract frame-level features and model time dependencies.

Therefore, Model A, B and C are defined as in Figures 5.2, combining the above-mentioned RGB-branches with the Pose-branch based on MLSTM-FCN.

All models concatenate the features provided by the two parallel branches, followed by the usual FC layer to transform the combined feature vector into a class-based binary vector. Therefore, the loss computes the error between predicted classes and the ground truth.

5.2.3 Training Modes

The aim of this section is to define two training modes for the Joint RGB-poses models. The goal is to make the most from the two available modal-

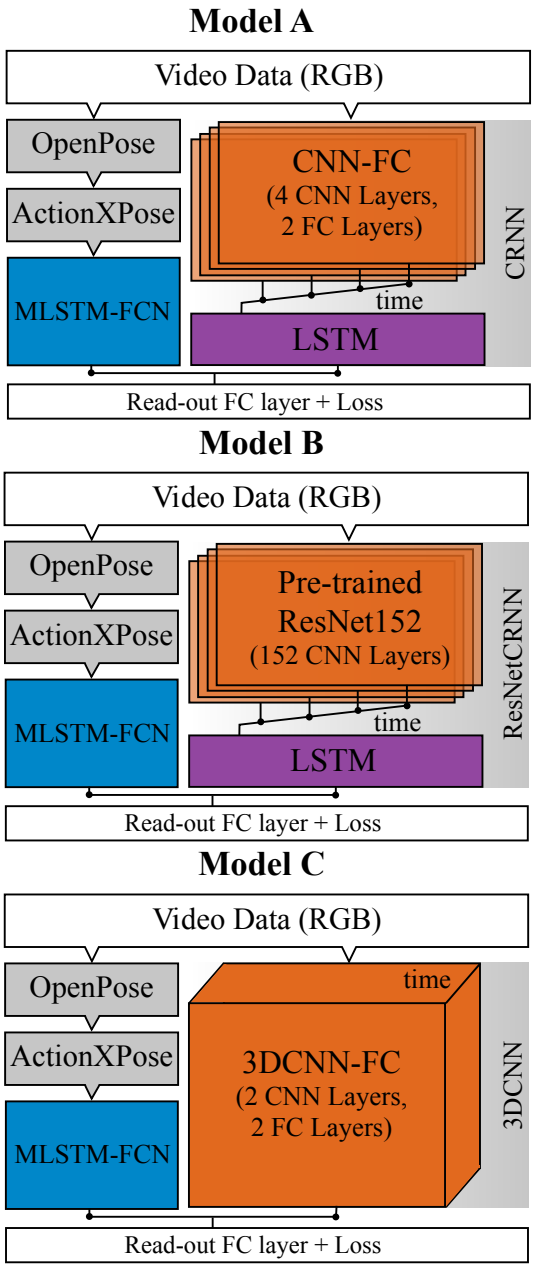


Figure 5.2: Proposed Models A, B, and C pipelines for joint RGB-poses HAR. Video RGB data is firstly pre-processed with OpenPose and ActionXPose to extract pose-related features to feed the poses-based learning branch performed by the MLSTM-FCN. In parallel, raw RGB data feeds the RGB-branch. Model A is based on a simple Convolutional Neural Network (CNN) structure, to separately extract visual features from each ROIs. Similarly, Model B exploits the same approach but using a pre-trained ResNet152 network. Thus, an additional LSTM layer is exploited to model temporal dependences between ROIs visual features vectors. As opposite, Model C RGB-branch is based on 3D Convolutional Neural Networks (3DCNN) and Fully Connected (FC) layers, to separately extract spatio-temporal visual features from each ROIs. For all models, resulting feature vectors extracted by both branches are concatenated and subsequently transformed by an FC layer, followed by the loss computation.

ities, i.e. RGB and pose.

Parallel Learning *(PL)*

In this mode, the two parallel learning branches are trained simultaneously. Nevertheless, the whole information I is available for the RGB-branch, while only P is available for the pose-branch. Therefore, while the training curve raises and the network weight are adjusted, some of the knowledge extracted by the RGB-branch might be the same potentially available in the Pose branch. This is because, while the pose-branch is learning how to exploit information in P, the RGB-branch might be faster in exploiting the same information. Therefore, in the subsequent learning iteration, the pose-branch might be discouraged to consider again information already exploited by the RGB-branch (Figure 5.3-Left). The result is that the whole, potentially, available knowledge in P might not be fully exploited.

Hierarchical Learning *(HL)*

To overcome the potential problem mentioned above, inspired by transfer learning [13], it is proposed to split the training phase into steps as follows:

a) *Network inizialisation*: random initialisation of the layers' weights;

b) *RGB-branch freezing*: RGB-branch layers cannot be optimised;

c) *Model training*: the training starts but the RGB-branch is kept frozen;

d) When the training curve starts overfitting the training stops and the best validation model is loaded;

e) *Unfreezing of the RGB-branch and freezing the pose-branch*: therefore, the pose-branch is kept as it is and we allow the RGB-branch to the updated. The FC layer is randomly re-initialised;

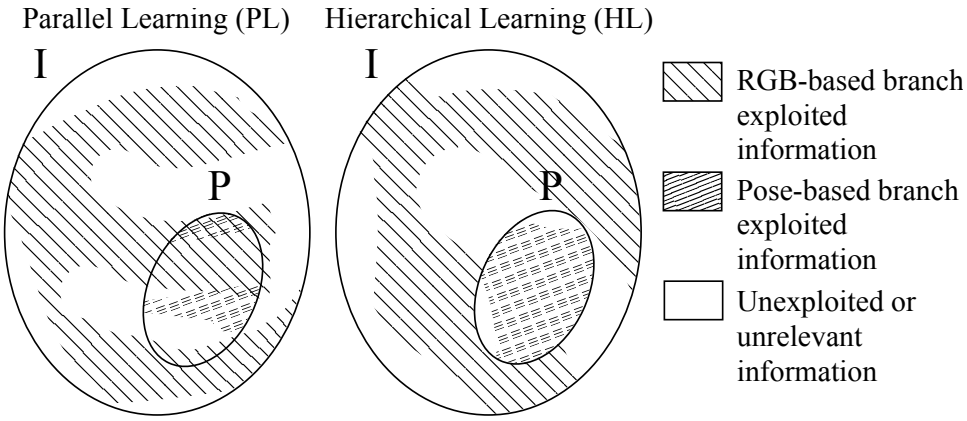


Figure 5.3: Training modes information coverage illustration. (Left) In the PL case, the RGB-branch might override P, which is the area of competence of the pose-branch, resulting in suboptimal usage of the two modalities. (Right) In the HL case, the PL potential problem is mitigated due to the separated usage of the two information sets P and $I \setminus P$.

f) *Model training*: the training starts again, but the pose-branch is no longer learnable;

g) When the training curve starts overfitting again, stop training and select the best validation model.

The idea behind HL is shown in Figure 5.3-Right. The HL training firstly takes the most from information P provided by pose data. Subsequently, once the pose-branch has solved the problem at its best, the RGB-branch is activated to explore information $I \setminus P$ and improve the obtained results. In practice, this strategy allows the RGB-branch to *only* solve the part of the problem that poses cannot solve, driven by the optimisation of the common loss function.

5.3 Proposed Combined HAR and HAAD

In this section, the strategy to combine HAR and HAAD is proposed, by exploiting the Models A, B and C defined in Section 5.2. The proposed, combined approach is *bi-level*, which means that it is able to detect two types of anomalies, i.e. abnormal body movements and abnormal object

positions.

5.3.1 Body Movements Based Combined HAR-HAAD

In this section, the models in Section 5.2.2 are exploited for combined Body Movements (BM) based combined HAR-HAAD. Let suppose that a joint RGB-models (A, B or C) has been trained for HAR on a pre-defined dataset (different than $\mathbb{D}$), with the unique restriction to consider k frames length action sequences. Let $\mathcal{M}_k$ be this model. Let also consider that RGB data associated with the action sequences consist only on the target's ROIs. Therefore, it is assumed that, to some extent, the model $\mathcal{M}_k$ can also perform HAR on unseen datasets. In particular, let us assume that $\mathbb{D}$ is the unseen dataset. Therefore, as shown in Figure 5.4, clips in $\mathbb{N}$ can be processed in the training phase to extract pose-branch features (PBB). Thus, PBB features are exploited to train an SVM one-class model for semi-supervised anomaly detection [144]. Similarly, $\mathbb{T}$ clips are processed in the testing phase to extract PBB features, to be injected into the trained SVM model, to determine whether the testing sequence should be considered as normal or abnormal. Since each PBB feature vector corresponds to a certain time window $[t_1, \ldots, t_k]$ for a given target ϕ, therefore $G_{\mathbf{s}}^{BM}([t_1, \ldots, t_k]|\phi)$ is computed. The superscript BM denotes that G is obtained considering poses-based body movements related features only. In principle, RGB-Based features (RBB) can also be exploited for the same task. Similarly, combinations between PBB and RBB, i.e. PBB+RBB, can be also exploited for HAAD. However, in Section 5.4.3, it is proven that PBB features outperforms both RBB and PBB+RBB features for HAAD.

It is worth mentioning that, regardless the human action labels included in the training of $\mathcal{M}_k$, $\mathcal{M}_k$ is always able to extract features from *any* and *unseen* k frames-length action sequence. If the unseen sequence resembles to one of the $\mathcal{M}_k$ training sequences, the resulting feature vector is such

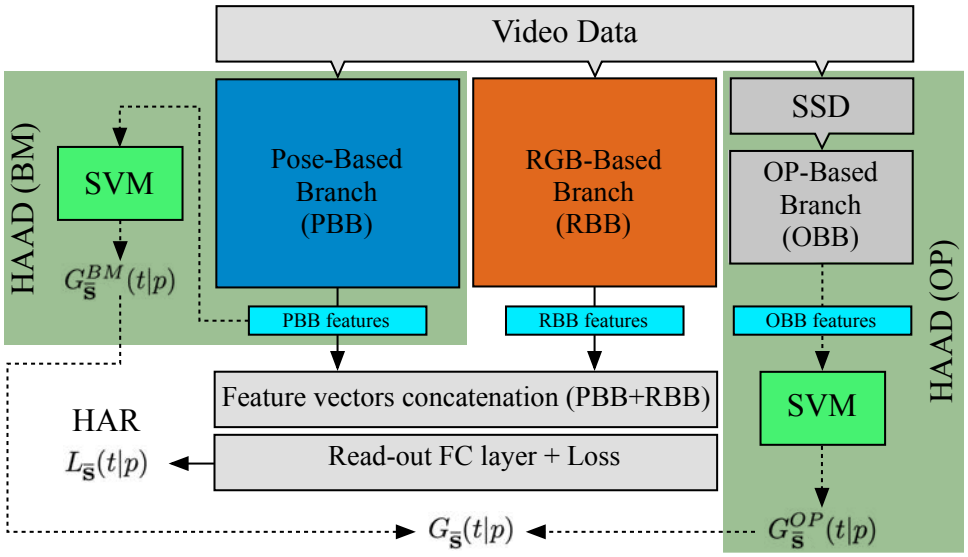


Figure 5.4: Proposed pipeline for bi-level, combined HAR and HAAD. Video data is processed by the pose-branch, RGB-based branch and the OP-based branch, obtaining, respectively, PBB, RBB and OBB features. PBB and RBB features are concatenated for joint RGB-poses HAR. In parallel, PBB features and OBB features feeds separately two SVM one-class model for anomaly detection. The final anomaly detection output is given by combining BM and OP based HAAD outputs.

that the read-out FC layer can reasonably provide the expected action label $L_{\bar{\mathbf{s}}}([t_1, \ldots, t_k]|\phi)$. As opposite, if the unseen sequence does not resemble training data for $\mathcal{M}_k$, the resulting action label will be not necessarily a correct description of the performed action. Nevertheless, the SVM model can still be trained on those features, regardless of the action-related knowledge they are carrying.

In this work, as first important approximation, it is assumed that posture-related HAAD is mostly independent from the object position. Therefore, in the next section, object position based HAAD is proposed as additional, independent level of HAAD.

5.3.2 Object Position Based HAAD

In this section, the goal is to perform anomaly detection based on the relative position between the target ϕ and the contextual Objects Position (OP). Let us assume that the dataset $\mathbb{D}$ contain a given set of w key objects O (objects

under monitoring, including the human target). For each given frame t, let

$$\begin{aligned} &h = 1 \;\longrightarrow\; (\phi_x, \phi_y) && \text{Human target} \\ &h = 2, \ldots, w \;\longrightarrow\; \{(x,y)_{h,1}, \ldots, (x,y)_{h,J_h}\} && \text{Other objects} \end{aligned} \tag{5.3.1}$$

be the positions for the objects in O, where J_h is the number of occurrences of the h-th object in O. Therefore, anomalies might come from the unexpected mutual positions between different objects in O. Thus, for each human target ϕ and for each frame t, multiple position-based $\mathbf{q}(t,\phi) \subset \mathbb{R}^{2w}$ feature vectors can be defined, one for each occurrence, as follows:

$$\begin{aligned} &\mathbf{q}(t,\phi) = \{[\phi_x, \phi_y, x_{2,j_2}, y_{2,j_2}, \ldots, x_{w,j_w}, y_{w,j_w}]\} \\ &\forall j_2 \in [1, \ldots, J_1] \;, \ldots, \; \forall j_w \in [1, \ldots, J_w] \end{aligned} \tag{5.3.2}$$

The advantage of this approach is that it allows straightforward generalization to multiple human targets and multiple occurrences of the same object type in the scene. Let the Object Position-based Branch (OBB) be the set of OP feature vectors, as in Figure 5.4, i.e. $\mathbf{q}(t,\phi) \subset \text{OBB}$ for a given frame t and target ϕ. Similarly to the previous section, a one-class SVM model can be trained by using normal features OBB extracted from $\mathbb{N}$ and tested OBB features extracted from $\mathbb{T}$. However, as opposed to the previous section where k frames where considered at a time, the abnormality estimation is based on *single* frames by using multiple feature vectors as follows:

$$G_{\tilde{\mathbf{s}}}^{OP}(t|\phi) = \begin{cases} 0 \;\text{ if }\; \text{SVM}(\mathbf{q}(t,\phi)) = 0 \;\;\forall \mathbf{q}(t,\phi) \\ 1 \;\text{ if }\; \exists\, \mathbf{q} \in Q(t,\phi) \;\; s.t. \;\; \text{SVM}(\mathbf{q}(t,\phi)) = 1 \end{cases} \tag{5.3.3}$$

where $\text{SVM}(\mathbf{q}(t,\phi))$ denotes the trained SVM model tested on the vectors $\mathbf{q}(t,\phi)$.

5.3.3 Bi-level BM-OP Activity Anomaly Detection

In Section 5.3.1, it has been assumed that BM and OP anomalies are independent. In this section, this assumption is exploited to combine anomaly detections $G_{\bar{\mathbf{s}}}^{BM}(t|\phi)$ and $G_{\bar{\mathbf{s}}}^{OP}(t|\phi)$. Therefore, the combination can be obtained by using a *Logic OR*, as follows:

$$G_{\bar{\mathbf{s}}}(t|\phi) \approx \begin{cases} 0 & \text{if} \quad G_{\bar{\mathbf{s}}}^{BM}(t|\phi) = 0 \ \wedge \ G_{\bar{\mathbf{s}}}^{OP}(t|\phi) = 0 \\ 1 & \text{otherwise} \end{cases} \tag{5.3.4}$$

where $G_{\bar{\mathbf{s}}}(t|\phi)$ is the aimed estimation of the anomaly detection ground truth for the $\bar{\mathbf{s}}$ video clip. See Figure 5.4 for a general overview of the proposed bi-level approach.

5.3.4 ROC curves for SVM one-class models

In this section, a simple solution to obtain ROC curves for SVM one-class models is proposed. To the best of the knowledge, there are no standard methods to provide ROC curves for SVM one-class based anomaly detection models. However, the ROC curve is a common metric for anomaly detection systems and allows effective comparisons between different models performance. A smooth (not binary) decision surface, e.g. based on a Logistic function, is required to compute ROC curves. However, it is impossible to fit a Logistic function by using only one-class data properly. Therefore, it is proposed to weaken this requirement by imposing the logistic function to be part of the anomaly detection model, fixing a-priori its parameters.

Let $s \in [-\infty, \infty]$ be the obtained SVM score for testing data. It is proposed to simply transform s by using a logistic function as follows:

$$f(s) = \frac{1}{1 + e^{-k_0(s-s_0)}} \tag{5.3.5}$$

where s_0 is the sigmoid midpoint and k_0 is the curve steepness. Samples

such that $s = 0$ lay on the edge of the SVM decision boundary. Conversely, samples such that $s > 0$ or $s < 0$ lay, respectively, inside or outside the normal region. Therefore, it is natural to set $s_0 = 0$, in order to obtain $f(s = 0) = 0.5$, $f(s > 0) > 0.5$ and $f(s < 0) < 0.5$. The parameter k is inversely proportional to the confidence of the model and can be fixed a-priori. The defined $f(s)$ represents the new score function which allows the ROC curve computation. This definition of $f(s)$ requires to experimentally set the parameter k_0. However, it does not include any learnable parameter and, once the parameter k_0 is fixed, it can be merged as part of the SVM model when ROC curves are required.

5.4 Simulations and Results

In this section, experiments for supporting the Joint RGB-poses methodology for HAR in Section 5.2 are reported in Sections 5.4.1. The bi-level combined HAR-HAAD methodology discussed in Section 5.3 is supported by experiments in Section 5.4.3.

5.4.1 Joint RGB-poses Simulations for HAR

UCF101 Dataset

In this section, Model A, B and C are tested on the popular UCF101 dataset for HAR. UCF101 include 101 action classes related to daily life activities. These actions are considerably determined by contextual information which are not carried by the body poses. For example, the *playing-cello* action is strongly determined by the played instrument RGB data. Therefore, despite the body pose data might be informative in some extent, RGB data is crucial to dispel any doubt. In Figure 5.5-(Top), four examples from UCF101 are depicted.

Since UCF101 shows multi-target video clips, it is likely that multiple

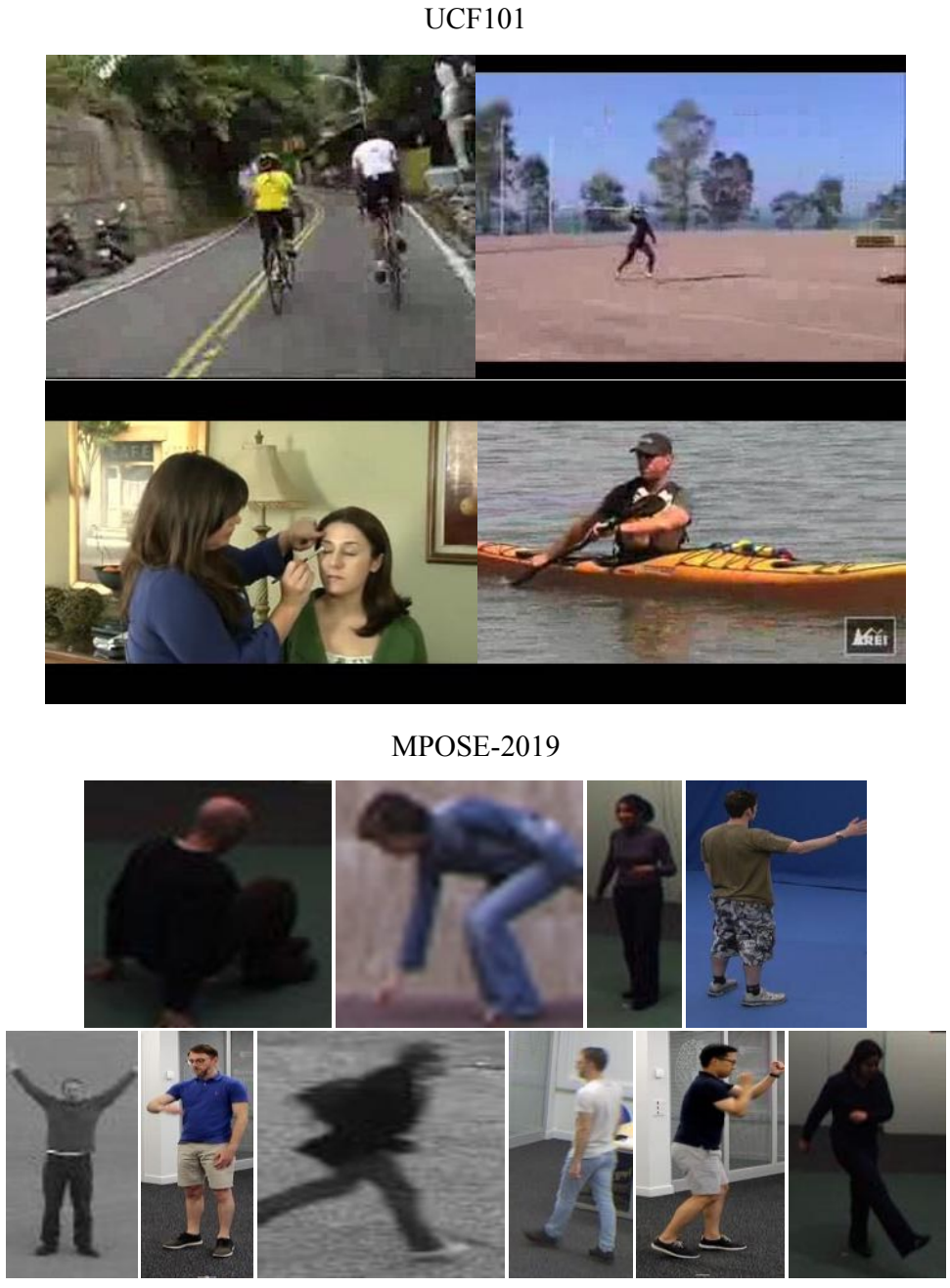


Figure 5.5: (Top) Examples from UCF101 dataset [31]. (Bottom) Examples from MPOSE-2019 dataset, which includes Weizmann [6], KTH [7], i3DPost [8], IXMAS [9] and the proposed ISLD datasets.

targets are detected in the same frame. However, this dataset includes numerous low-quality and low-resolution video samples where OpenPose fails, as already discussed in detail in Section 4.4.2. Due to these limitations, in this section the ActionXPose step is skipped. Therefore, in the pose-branch, OpenPose directly input detected poses to the MLSTM-FCN, regardless the pose landmarks quality. Furthermore, this dataset includes single and multi-target video clips. However, each video clip is provided with a single label that refers to the whole clip rather than to a label for each single target. Therefore, pose data is passed through the pose-branch frame-by-frame alongside the target identity as additional input feature, to allow MLSTM-FCN do potentially discriminate between different target data. Moreover, for the same reason, the whole frame is passed through the RGB-based branch.

In Table 5.1, Model A, B and C results are compared with those obtained by the MLSTM-FCN (pose-branch alone) and the CRNN, ResNetCRNN and 3DCNN (RGB-based branch alone). Since MLSTM-FCN reaches 60.9% accuracy, the body poses knowledge is not enough to effectively describe the actions. On the other hand, despite RGB data is expected to be fully informative, the knowledge extracted from RGB-based models varies depending on the considered structure. In particular, RGB-based models reach respectively 56.1%, 79.9% and 50.8%. As expected, despite full information is available, the knowledge extracted depends on the model. Nevertheless, the Joint RGB-poses approaches always outperform both the RGB-based models and the poses-based model, demonstrating that the proposed multimodality is effective and beneficial. Moreover, the PL training mode is outperformed by the HL training mode, which supports the arguments in Section 5.2.3.

MPOSE-2019 Dataset

In this section, Model A, B and C are tested on a challenging dataset called MPOSE-2019. MPOSE-2019 is the extension of the MPOSE dataset, which

Table 5.1: UCF101 dataset (HAR) performance. Accuracies (Acc) for different and different training modes, i.e. PL and HL, depending on the data modalities.

	Modalities	Acc	Acc (PL)	Acc (HL)
MLSTM-FCN	Poses	60.9%	n/a	n/a
CRNN	RGB	56.1%	n/a	n/a
Model A	RGB-Poses	n/a	63.0%	78.1%
ResNetCRNN	RGB	79.9%	n/a	n/a
Model B	RGB-Poses	n/a	84.4%	91.1%
3DCNN	RGB	50.8%	n/a	n/a
Model C	RGB-Poses	n/a	61.2%	67.12%

has been presented in Chapter 4. In particular, the ISLD dataset, which was part of MPOSE, is extended to ISLD-2019. This extension include additional sequences for some actions, i.e. *walking*, *hand-waving*, *boxing*, *pointing*, *bending*, *hands-clapping*, *hands-waving* and *running*, to extend the viewpoints variability. Additional sequences were recorded in the Intelligent Sensing Lab. Thus, in this chapter, MPOSE-2019 is defined by fusing Weizmann, IXMAS, i3DPost, KTH and ISLD-2019. In Figure 5.6, a summary of the action composition of MPOSE-2019 is provided.

As discussed in Chapter 4 for MPOSE, MPOSE-2019 represents the best scenario where ActionXPose can be effectively used, since the target body is potentially fully visible. Therefore, as opposed to the previous section, in this section ActionXPose is used to transform pose data provided by OpenPose into meaningful temporal-motion features, as discussed in Section 5.2.1.

It is worth underlining that actions into MPOSE-2019 are mainly posture-related and unrelated to the background. For the purpose of this chapter, background information have been further decreased by considering only target's ROIs, neglecting the rest of the frame background. Figure 5.5-Bottom shows some examples of ROIs provided by MPOSE-2019. Therefore, as opposed to the previous section, in this section target's ROIs are provided to the RGB-branch in Model A, B and C.

The MPOSE-2019 results reported in Table 5.2 follows the same trend

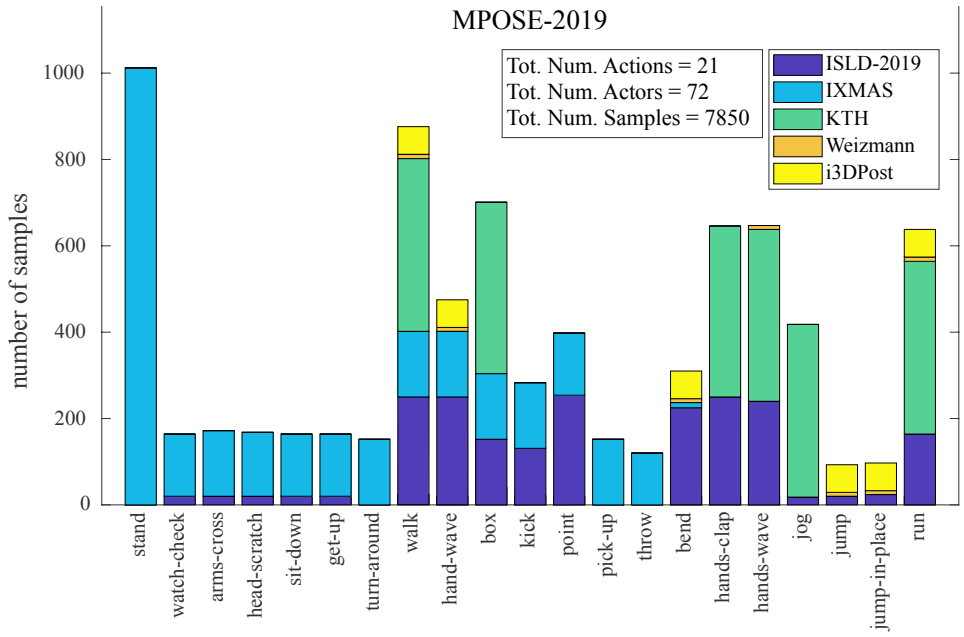


Figure 5.6: MPOSE-2019 actions statistics. MPOSE-2019 contains 7850 video clips collected from famous posture-based HAR datasets. Video clips contains 21 posture-related actions performed by 72 human targets.

Table 5.2: MPOSE-2019 dataset (HAR) performance. Baseline models, i.e. the poses-based MLSTM-FCN and the RGB-based models, are compared with the proposed Joint RGB-poses models, i.e. Model A, B and C.

	Modalities	Acc	Acc (PL)	Acc (HL)
MLSTM-FCN	Poses	90.3%	n/a	n/a
CRNN	RGB	90.4%	n/a	n/a
Model A	RGB-Poses	n/a	90.8%	94.9%
ResNetCRNN	RGB	80.8%	n/a	n/a
Model B	RGB-Poses	n/a	89.4%	95.8%
3DCNN	RGB	77.3%	n/a	n/a
Model C	RGB-Poses	n/a	76.3%	90.4%

already seen in Section 5.4.1. In fact, the poses-based and RGB-based models are outperformed by the Joint RGB-poses models, especially in the case of HL training mode. This results further confirm the multimodal approach validity.

Activation Maps for HAR

The goal of this section is to explore how CNN activation maps [145] changes in the case of Joint RGB-poses HAR. To this purpose, for simplicity, let us consider models CRNN and Model A, trained on MPOSE-2019, since the RGB branch consists of only four CNN layers. In Figure 5.7, each layer activation map is depicted in the CRNN case, i.e. no pose data was available during training, and Model A, i.e. pose data was available during training. In particular, Model A has been trained with the HL training mode proposed in Section 5.2.3. Therefore, it is expected that Model A's RGB-branch focuses on different areas, since poses are providing information that the RGB-branch can avoid to extract. This expectation is confirmed by a visual inspection of the obtained activation maps.

In Figure 5.7-(a,b,c), three video clips are passed through CRNN and Model A and the activation maps for each layer is shown (first video clip frame only). The CRNN is clearly focusing on the most of the ROI (see ALL column), revealing that contextual information as well as target details are needed to take the final decision. On the other hand, Model A's RGB-branch is clearly only using target fine details to take the final decision, since complementary information is provided by the poses. These examples suggest that when the poses-based branch undoubtedly detect the right action, the RGB-branch is requested to focus on contextual details to support the other branch or at least on regions that to not contradict the pose-branch.

Nevertheless, it is interesting to note that, for some examples, the above-mentioned behaviour is reversed. As shown in the examples in Figure 5.7-(d,e,f), the CRNN is clearly relying on finer target details, while the Model A considers more contextual information. This is again due to the contribution of the poses. However, as opposed to the previous case, these examples seems to suggest that when the pose-branch information is unreliable, the RGB-based branch is encouraged to focus on any useful detail that allows the

cost function to drop. Alternatively, if the pose-branch is reliable, the RGB-branch is encouraged to focus on regions that confirm (or not contradict) the pose-branch information. In both cases, the multimodal approach activation maps considerably change compared to the RGB-based case, supporting the theoretical ideas discussed in Section 5.2.2.

5.4.2 BMbD, M-BMbD and JBMOPbD Datasets

In literature, there is a multitude of datasets for video-based anomaly detection. UCF-Crime [108] is a huge and challenging dataset which includes multiple abnormal instances from several and different semantic events. However, the proposed human actions abnormal classes are limited and not related to posture. Therefore, this dataset is not suitable as a benchmark for a poses-based method. Other datasets such as Hockey Fight, Subway, UMN, Violence in Crowds (VIC), Violence in Movies (VIM) and BEHAVE [31, 48, 117, 146] are also not suitable for highlighting the efficacy of the proposed pose-driven HAR and HAAD. The first limitation regards the lack of explicit human posture based anomalies. Targets bodies are mostly out from the camera field of view or, even if fully visible, the targets are not performing abnormal actions in terms of postures. The second limitation is related to objects. Indeed, visible objects positions do not explicitly define an anomaly detection problem which can be used to show the efficacy of the proposed method and the limitations of state-of-the-art approaches. For these reasons, novel, non-collaborative, challenging datasets, BMbD, M-BMbD and JBMOPbD were recorded in the Intelligent Sensing Lab, focusing on posture-related and object position-related anomalies.

The recording set up includes two standard RGB cameras mounted in a fixed position on two top-corner of the Intelligent Sensing Lab. The dataset includes spontaneous actions performed by up to 3 actors, wearing different clothes to include appearance change challenges. Additionally, three key

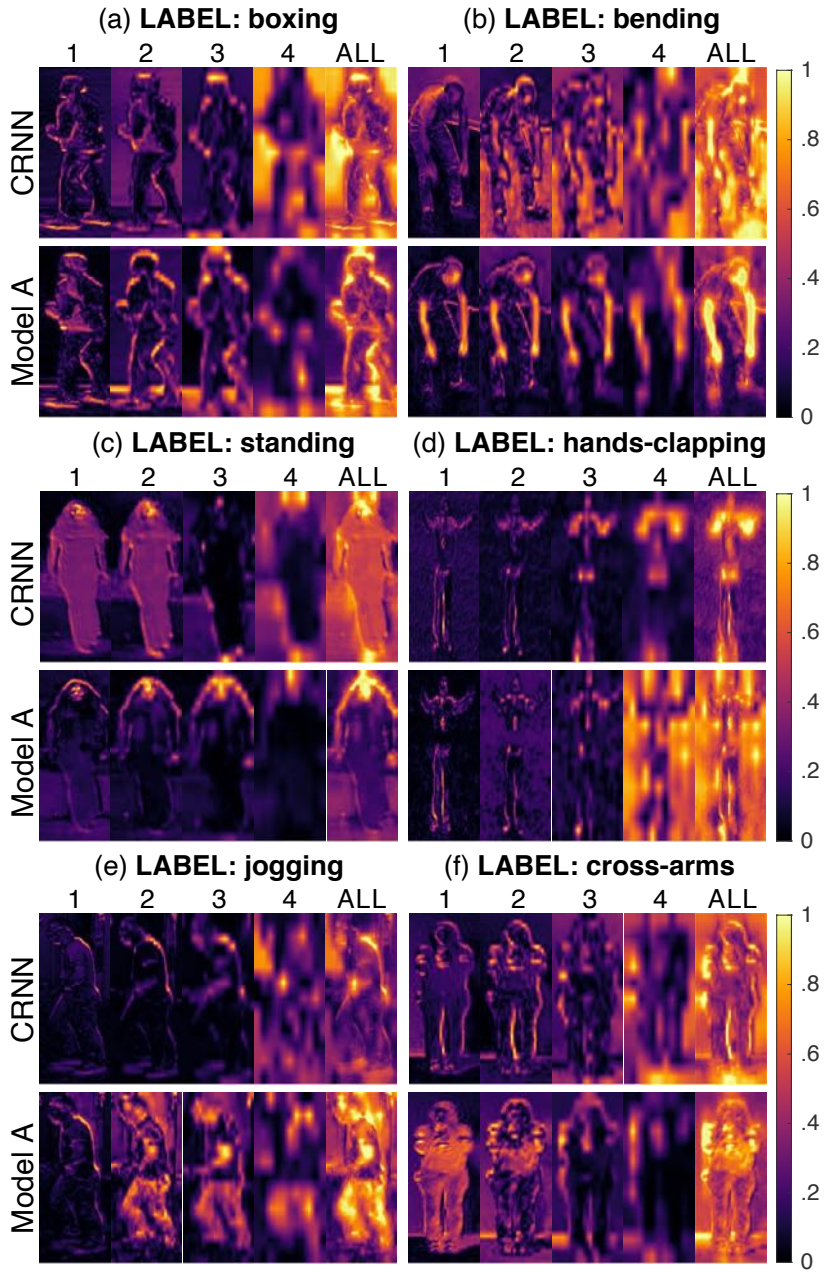


Figure 5.7: Comparison between first-frame activation maps for CRNN and Model A. The four CNN levels activation maps are depicted (1, 2, 3 and 4), alongside the cumulative activation map (ALL) obtained by summing up the others. It is evident that when pose information are available (Model A) the RGB-branch is encouraged to focus on different areas than those considered when pose information is not available (CRNN). In cases (a), (b) and (c), CRNN is more activated by contextual information, while Model A RGB-branch is more activated by finer target details. In the (d), (e) and (f) cases, CRNN is more activated by finer target details, while the Model A RGB-branch is more activated by contextual information and/or target details.

objects types were considered, i.e. human target, chair and a bike, to the purpose of OP based anomaly detection. The human targets are also free to interact with other environmental objects, such as drawers, tables, pc monitors, and other human targets. Video data within the dataset is divided into *normal* and *testing*, to allow semi-supervised anomaly detection, as mentioned in Section 1.3.1. Regarding the performed action anomalies, they consist of: 1) unexpected body movements, not necessarily among a pre-defined set of actions; 2) key object position anomalies. In particular, regarding JBMOPbD normal dataset, the key objects position was constrained to be within rectangular marks on the floor. As opposed, in the JBMOPbD testing subset, the targets were free to move the objects in any position.

Table 5.3 summarizes the abnormal semantic domains for each proposed dataset, to independently/jointly evaluate BM, OP, single and multi-target performance. The ground truth for testing videos in ISLD-A has been manu-

Table 5.3: ISLD-A splits features summary. Total training/testing time is reported. HAAD targeted levels BM/OP are checked accordingly. Single, multi and not considered (n/c) human target mode is also reported.

Split	Training Time	Testing Time	BM	OP	Target
BMbD	21m:34s	19m:42s	✓	n/a	single
M-BMbD	21m:34s	05m:28s	✓	n/a	multi
JBMOPbD	21m:38s	12m:44s	✓	✓	single

ally set for HAAD. Therefore, for each tracked human target, a binary label normal/abnormal has been fixed for each target bounding box, frame-by-frame. Regarding HAR, no ground truth has been set, since the performed actions were spontaneous and difficult to be always clearly classified according to a pre-defined set of actions.

As evaluation metric for HAAD, it is suggested to consider the overall frame-level average accuracy, which measures the following ratio

$$\frac{1}{N}\sum_{t=1}^{N}\frac{\mathbf{TP}(t)+\mathbf{TN}(t)}{\mathbf{P}(t)+\mathbf{N}(t)} \tag{5.4.1}$$

where $\mathbf{TP}(t)$ and $\mathbf{TN}(t)$ represent respectively the true positive (abnormal) and true negative (normal) detections at frame t, $\mathbf{P}(t)$ and $\mathbf{N}(t)$ represent the real positive (abnormal) and real negative (normal) detections at frame t, for all available frames N. When multiple human targets are available on the same frame t, thus $\mathbf{TP}(t)$, $\mathbf{TN}(t)$, $\mathbf{P}(t)$ and $\mathbf{T}(t)$ are cumulative over visible targets. As additional metric, ROC curves are considered to allow effective state-of-the-art comparisons.

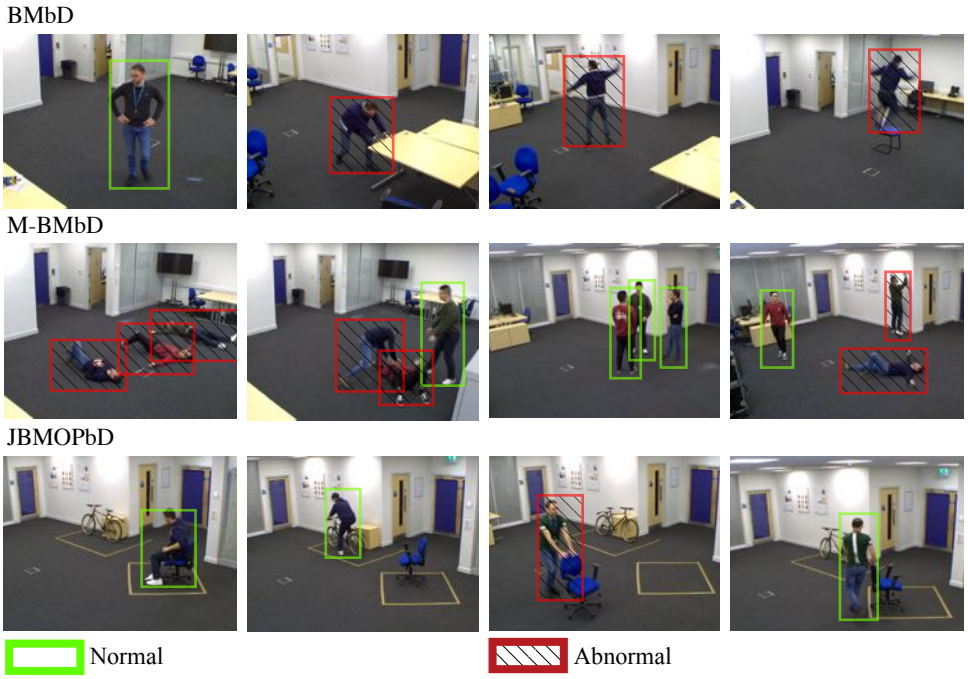


Figure 5.8: BMbD, M-BMbD and JBMOPbD datasets examples. Green bounding boxes depict the normal events, while the strikethrough red bounding boxes represent abnormal events.

5.4.3 Bi-level Combined HAR-HAAD Evaluations

As mentioned in Section 5.3, the combined HAR-HAAD model trains its HAR abilities on a different dataset. MPOSE-2019 is used to this purpose, since it best fits the criteria for combined HAR-HAAD on the novel datasets introduced in the previous section. Moreover, as seen in Section 5.4.1, Model B achieves the best HAR results on MPOSE-2019. Therefore, Model B

is chosen as main architecture to test the bi-level combined HAR-HAAD strategy.

State-of-the-art Comparisons

In this section, comparisons between Model B, ResNetCRNN and official implementation of available algorithms [109, 110] for video-based anomaly detection is provided. The chosen benchmarks are BMbD and JBMOPbD splits. [109, 110] provide ROC curves as standard performance evaluation metric. The best ROC curve for the proposed SVM one-class models is computed as proposed in Section 5.3.4 by experimentally fixing $k_0 = .2$.

Obtained results are depicted in Figure 5.9. The graph shows the ROC curves for Model B, ResNetCRNN and [109, 110]. Regarding Model B and ResNetCRNN, curves for $\alpha = [.01, .04, .07, .1, .13, .16, .19, .22]$ and $k_0 = [.1, .2, .3, .4, .5, .6, .7, .8]$ are drawn to show the performance impact of different parameters. In the case of JBMOPbD, BM and OP based HAAD are selectively enabled/disabled, to highlight the importance of both levels of analysis.

Overall, the graph shows that the proposed Model B greatly improve performance over both ResNetCRNN and the state-of-the-art models based on autoencoders. In particular, autoencoders failure is likely due to the fact that the abnormality is not merely related to strong changes on the RGB content. In fact, when the human target starts performing an abnormal action, the target appearance remains mostly preserved, preventing the autoencoder to fail in reconstructing the RGB data. In contrast, the proposed approach relies on contextual and pre-trained knowledge which is effectively transferred to the HAAD problem.

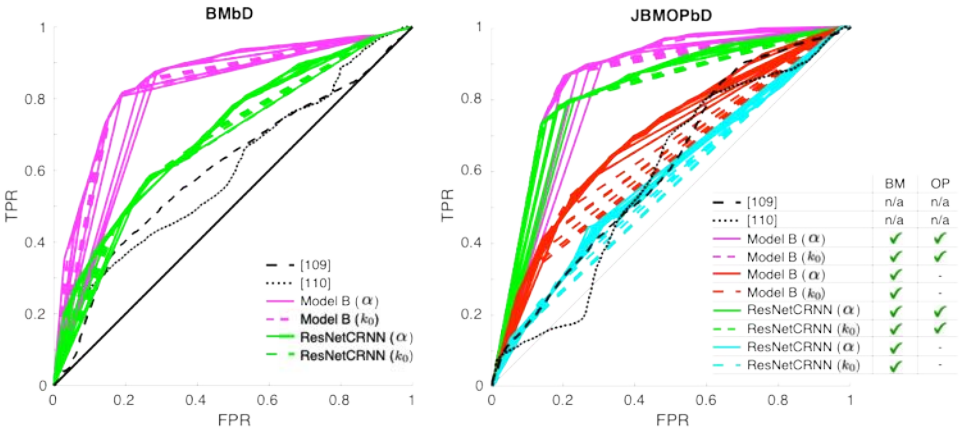


Figure 5.9: Model B, ResNetCRNN and state-of-the-art methods performance comparison. (Left) ROC curves for the BMbD dataset obtained by varying parameter $\alpha = [.01, .04, .07, .1, .13, .16, .19, .22]$ and $k_0 = [.1, .2, .3, .4, .5, .6, .7, .8]$. (Right) ROC curves for the BMbD dataset obtained by varying parameter $\alpha = [.01, .04, .07, .1, .13, .16, .19, .22]$ and $k_0 = [.1, .2, .3, .4, .5, .6, .7, .8]$, and enabling/disabling BM/OP anomaly detection.

HAAD Performance Study

In this section, details on how to train the combined HAR-HAAD model are provided. Let Model B be the chosen architecture, denoted as $\mathcal{M}_k$, and trained on MPOSE-2019 by using the HL training mode. The knowledge learned from MPOSE-2019 for HAR is exploited to perform HAAD on other datasets, e.g. BMbD, M-BMbD and JBMOPbD. As discussed in Section 5.3.1, in this section it is imposed MPOSE-2019 training/validation data do be randomly cropped by using an experimentally set 30-frames time window, i.e. $k = 30$. Therefore, at each training epoch, training samples are randomly cropped to increase generalisation. As opposed, validation samples are randomly cropped only at the beginning of the training, to keep stable validation data over different epochs and obtain consistent improvements for the validation curve. Figure 5.10-(a) shows the validation curve obtained during the pose-branch HAR training. Only epochs $\mathcal{E} \subset \{1, \ldots, 100\}$ where the validation score increased are reported on the x-axis. The status of the network has been saved at each epoch $e \in \mathcal{E}$, progressively obtaining mod-

els $\mathcal{M}_{30}^{e}$, for all $e \in \mathcal{E}$. Therefore, for all $\mathcal{M}_{30}^{e}$, the HAAD corresponding accuracy on the HAAD datasets are reported on Figure 5.10-b)-e), for all dataset splits. Since BM related HAAD is based on a SVM one-class model which requires to set an additional parameter α, i.e. the *outlier fraction*, the accuracy obtained for $\alpha = \{.01, .04, .07, .10, .13, .16, .19, .22\}$ is reported. Regarding the outlier fraction parameter β required for OP based HAAD, it has been experimentally set to $\beta = .01$.

It can be seen that optimum HAAD performance are not necessarily reached in correspondence to the best epoch in terms of HAR performance. In fact, for BMbD and M-BMbD, the peak in at, respectively, the 23th and 24th epochs. For the JBMPObD, the peak in on the 17th epoch when OP based HAAD is not considered and on the 18th epoch when OP based HAAD is considered. This surprising behaviour might be due to the concurrence of two major causes. First, the BM features are not designed to optimise the SVM performance but to optimise the HAR-based loss function. Therefore, while the HAR performance follows an increasing monotonic curve, the HAAD performance might be not necessarily monotonic. Second, there might be a slightly overfitting effect, due to the chosen training strategy which randomly crops training sequences. In fact, as the HAR validation score on MPOSE-2019 increases, the sequence random cropping might force completely different sequences to have the same output in terms of features. For example, data that resemble a *standing* action often surrounds other actions in the same sequence. Therefore, randomly cropping the sequence around it forces data resembling the *standing* action to be labelled differently. This strategy improves generalisation for HAR, but might reduces the HAAD features sensitivity.

Regarding α, as expected, the parameter plays a role in optimising performance and needs to be chosen according to the tested dataset split for best performance.

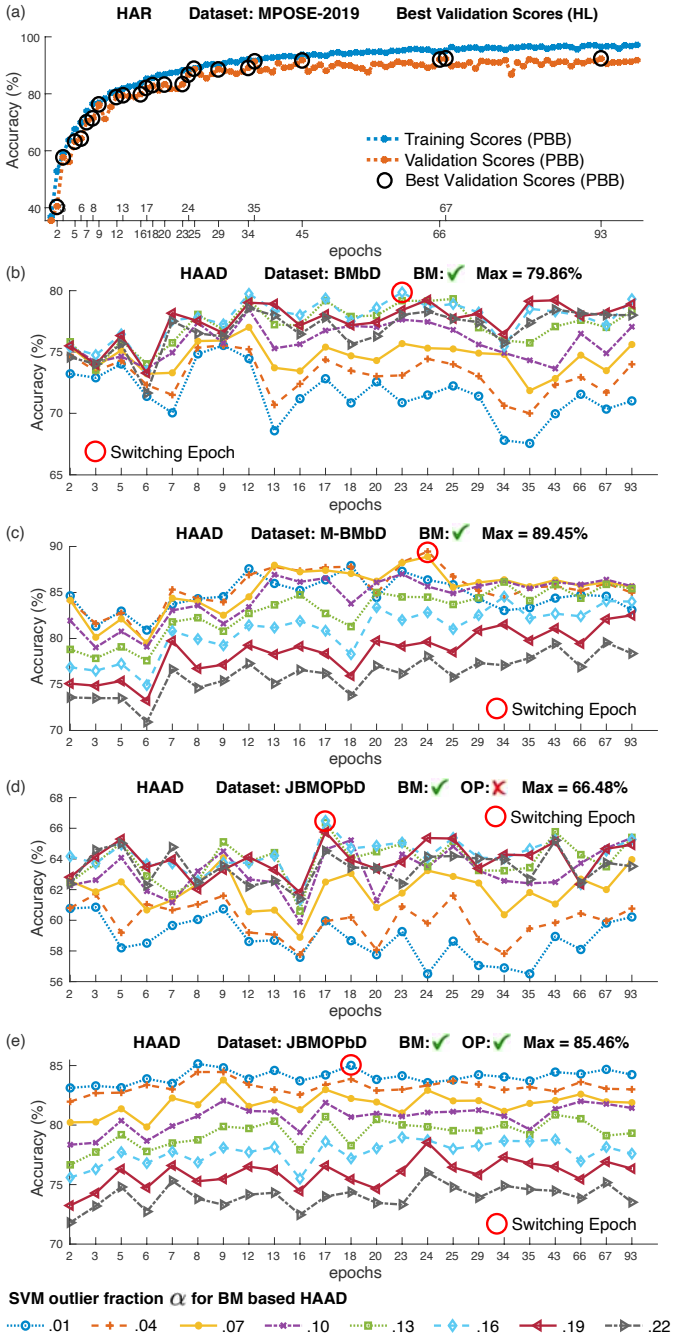


Figure 5.10: Cross-datasets (MPOSE-2019 to ISLD-A) HAAD and HAR epoch-based study for switching epoch (○) research (a) Model B pose-branch best validation scores (HL mode). (b) Corresponding HAAD accuracy for BMbD. (c) Corresponding HAAD accuracy for M-BMbD. (d) Corresponding HAAD accuracy for JBMOPbD, in the case of disabled OP based HAAD. (e) Corresponding HAAD accuracy for JBMOPbD, in the case of enabled OP based HAAD.

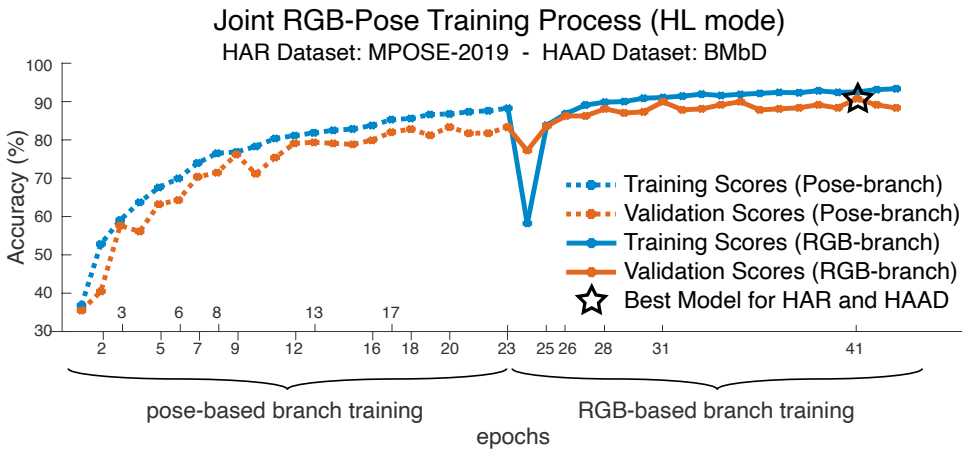


Figure 5.11: Joint RGB-poses Training Process optimised for HAAD for the BMbD dataset BMbD.

So far, the epoch e which optimises the HAAD performance using the pose-branch trained for HAR has been found. Therefore, in order to perform the RGB-branch training by using the HL training mode, the best epoch $\bar{e}$ in terms of HAAD is set as *switching epoch*, i.e. the epoch where the RGB-based branch training starts and the pose-branch training ends. An example of the resulting training/validation curve obtained for the HL training mode is shown in Figure 5.11. The HL mode ensures that the pose-branch is frozen on the state $\mathcal{M}_{30}^{\bar{e}}$, therefore HAAD performance are preserved. Since the pose-branch status in $\mathcal{M}_{30}^{\bar{e}}$ might not be optimal for HAR, the RGB-branch training is used to compensate HAR performance.

In summary, when the HL training is complete, the pose-branch is able to extract features which are optimal for HAAD, suboptimal for HAR and compensated by RGB-based features for best HAR performance.

HAAD Ablation Study

In this section, the ablation study for HAAD is provided. This study aims to highlight the contribution of different features to the overall HAAD accuracy. In particular, all relevant combinations of features are systematically exploited following the methodology presented in Section 5.4.3. Therefore,

for each dataset, combinations of pose-branch features, i.e. PBB, and RGB-based branch features, i.e. RBB, are considered for the BM based HAAD. Moreover, for JBMOPbD dataset, BM based features are considered alongside OP based features, i.e. OBB, in all possible setups.

The best accuracy over epochs $\mathcal{E}$ is reported in Table 5.4 for each case, including relative α and β parameters.

As a useful comparison, in this study the ResNetCRNN performance as HAAD features extraction method is also reported. To this purpose, ResNetCRNN has been trained on MPOSE-2019 using the same 30-frames random time cropping already used for the Model B training. The obtained training process is depicted in Figure 5.12. Following the same approach already used for the Model B, several network status are saved by using the best validation scores. Therefore, resulting networks are used for HAAD features extraction.

The overall conclusion of this study is that RGB related features (RBB) are always worsening the HAAD results and should not be considered, supporting the study reported in Section 5.4.3. On the other hand, considering OP related features (OBB) is always beneficial, as expected, since the JBMOPbD dataset includes significant anomalies from the unexpected key object positions.

5.5 Critical Analysis

The Objective 7 introduced in Section 1.3 has been achieved by proposing Model A, B and C for joint RGB-poses HAR, which represents a generalisation of pose-based HAR to joint RGB-pose based HAR. In particular, it is shown that HAR performance can be improved by using multimodal pose-driven approaches instead of single RGB-based or poses-based deep learning.

Moreover, Objectives 8, 9 and 10 have been also achieved, by proposing

Table 5.4: HAAD Performance Ablation Study. Model B best accuracy and α parameter are reported for relevant combinations features, i.e. PBB, RBB and OBB for HAAD. Results are compared with those obtained by the model ResNetCRNN, which is the precursor of the Model B.

BMbD	**Features**		
Model	**BM**	**Best Accuracy**	α
Model B	PBB	**79.86%**	.16
	RBB	57.84%	.13
	PBB+RBB	78.06%	.16
ResNetCRNN	RBB	66.68%	.16

M-BMbD	**Features**		
Model	**BM**	**Best Accuracy**	α
Model B	PBB	**89.45%**	.04
	RBB	74.10%	.04
	PBB+RBB	88.50	.07
ResNetCRNN	RBB	86.17%	.07

JBMOPbD	**Features**				
Model	**BM**	**OP**	**Best Accuracy**	α	β
Model B	n/a	OBB	81.80%	n/a	.01
	PBB	n/a	66.48%	.16	n/a
	PBB	OBB	**85.46%**	.01	.01
	RBB	n/a	56.41%	.19	n/a
	RBB	OBB	81.46%	.01	.01
	PBB+RBB	n/a	65.20%	.13	n/a
	PBB+RBB	OBB	85.11%	.04	.01
ResNetCRNN	RBB	n/a	57.31%	.13	n/a
	RBB	OBB	82.76%	.01	.01

a combined HAR and HAAD approach based on SVM models to simultaneously output: 1) body-movements anomaly detection labels; 2) object position anomaly detection labels; 3) action recognition labels for single and multi-target video data.

In this chapter, it has been also demonstrated that state-of-the-art approaches based on autoencoders for semi-supervised anomaly detection are not capable to effectively detect human posture-related anomalies, and they fail in detecting objects positions related anomalies. In contrast, the proposed method greatly outperforms the state-of-the-art, as the ROC curve comparison in Figure 5.9 demonstrates.

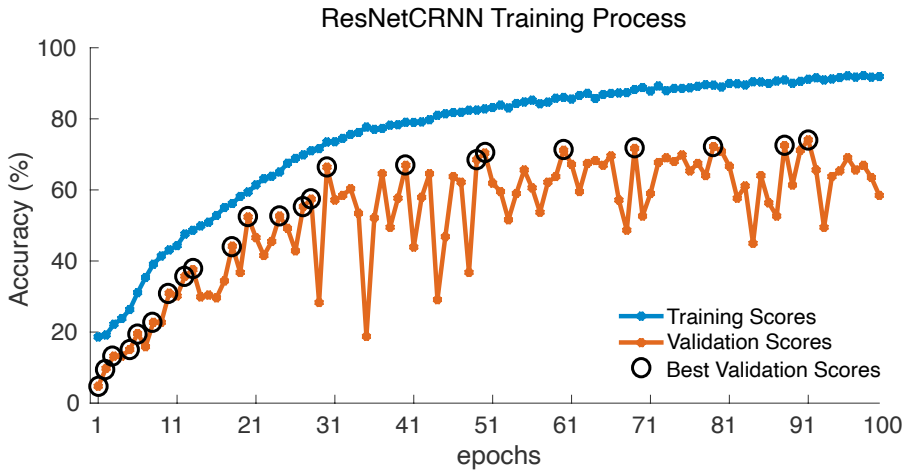


Figure 5.12: Training process for ResNetCRNN based on MPOSE-2019 using the 30-frames random time windows. The best validation network status is saved for HAAD features extraction. Best results obtained over those status are reported in Table 5.4.

Despite the achievements discussed in this chapter, some points can be raised, which set up possible future works.

The first point regards the state-of-the-art network I3D [19]. In [19], it has been shown that massive pre-training on huge datasets and multimodal video processing (RGB + optical flows) is beneficial for HAR. Therefore, in principle, I3D can be used as a standard for the RGB-branch, in place of the more general models considered in this thesis, i.e. CRNN, ResNetCRNN and 3DCNN. This can be done as part of future development, where the video processing can include RGB, optical flows and human poses in a three-way multimodal approach.

The second point regards the link between HAR and HAAD presented in this chapter. It has been shown that HAR performance growth during training is not necessarily linked with the HAAD performance growth. In other words, optimising HAR performance does not necessarily improve HAAD performance accordingly. This limitation is due to the fact that the HAR and HAAD model optimisations are not linked by a common loss function. Future work can further investigate this interesting point, proposing novel

fusion techniques to jointly optimise the model for the two distinct tasks.

The third and last point regards the fact that the proposed datasets, i.e. BMbD, M-BMbD and JBMOPbD, are not provided with HAR ground truth. As discussed in Section 5.4.2, it was difficult to categorise the performed action according to any predefined set of labels, particularly the action labels included in MPOSE-2019. This lack of ground truth made it impossible to explicitly quantify the advantage of joint RGB-poses HAR on these datasets over the unimodal HAR. However, it is possible to compare the results of RGB-based and joint RGB-poses based HAR on BMbD, M-BMbD and JBMOPbD, by visioning the output videos. By making this qualitative comparison, the advantages of the proposed method over the RGB-based one are evident. This result suggests that using pose-driven data is beneficial with respect to only using RGB data, i.e. target ROIs, to the purpose of more accurate HAR.

5.6 Chapter Summary

In this chapter, pose-driven approaches have been extensively studied in the context of HAR and HAAD. First, it is shown that joint RGB-poses approaches can improve RGB-based and poses-based methods efficacy for HAR. Second, novel datasets for posture-based, multi-target, non-cooperative HAAD are proposed. It is shown that autoencoder-based state-of-the-art method are suboptimal on the proposed datasets. Therefore, a novel and more effective solutions was proposed, which is based on joint RGB-posed networks. The proposed method is designed to simultaneously perform HAR and HAAD.

Chapter 6

CONCLUSIONS AND FUTURE WORK

Video-based Human Action Recognition (HAR) and Human Activity Anomaly detection (HAAD) offer great challenges in several applications domains. In this thesis, the posture-based HAR and HAAD have been considered. The goal was to achieve robust and computationally-efficient models which can be potentially used in non-cooperative, multiple human monitoring and surveillance. To achieve this goal, in Section 1.3, ten interlinked objectives were identified. Therefore, in Chapters 3, 4 and 5, the objectives were successfully addressed. Overall, this thesis contributed algorithms in the fields of silhouette-based HAR, pose-based HAR and pose-driven HAR and HAAD. Moreover, to support the experimental evaluation, novel posture-related, non-cooperative and multi-target datasets were also contributed, namely ISLD-2018, ISLD, ISLD-2019, BMbD, M-BMbD and JBMOPbD. Further and extensive evaluation was also conducted on publicly available and popular datasets.

In Chapter 3, the posture-related HAR problem was addressed by contributing new frameworks for silhouetted-based HAR. A popular method proposed by Weinland et al. [26] based on 3D-HOG and background subtraction was successfully replicated to serve as a baseline for further development. Thus, the limitations of the baseline were highlighted. Therefore,

the baseline limitations inspired the development of new algorithms, namely Framework A and B, for silhouette-based 3D-HOG based HAR. The proposed frameworks not only outperformed the baseline in terms of accuracy on public datasets but also overcame the baseline limitations. Despite these achievements, questions related to the background subtraction method were raised. The conducted experiment on ISLD-2018 highlighted that, in real-world recordings, the state-of-the-art ViBE algorithm for background subtraction requires to be intensively tuned. Moreover, it must be supported by an additional human detector, to distinguish the human targets from other objects and artefacts such as *ghosts*. Furthermore, the minimum computational effort which was possible to achieve was not encouraging. Overall, despite the obtained achievements, it was concluded that silhouette-based HAR, in the form discussed in this thesis, is not as promising as originally wished to achieve the goals of this thesis. However, this study allowed to acquire useful competencies in HAR and imagine other potentially disruptive solutions.

In Chapter 4, the expertise developed in the previous chapter was capitalised by driving the focus from silhouette-based HAR to pose-based HAR. The key step was to cast aside ViBE and prefer OpenPose as a posture-based data collector. The advantages of OpenPose with respect to ViBE are: 1) OpenPose performs human detection and posture-related data collection in one shot; 2) OpenPose consistently provides human-poses despite the considered dataset; 3) OpenPose does not require any intense parameter selection. Therefore, a novel algorithm was proposed to perform robust HAR on the basis of 2D human-poses. After extensive performance evaluation, the proposed algorithm, named ActionXPose, demonstrated to be as effective as those based on human-silhouettes in terms of accuracy. Moreover, the proposed solutions for dealing with body occlusions made ActionXPose more robust compared to the baselines. Overall, ActionXPose achieved performance

among the state-of-the-art over the tested datasets. Its generality also allows performing cross-datasets experiments. The definition of ActionXPose made a considerable step forward towards the goal of this thesis. In fact, despite the OpenPose limitations on in-the-wild scenarios, ActionXPose showed high robustness and computational-efficiency in several, posed, multi-viewpoint and publicly available datasets. However, it was still necessary to test it on non-cooperative scenarios. Moreover, since human-poses lack of semantic information other than the body limbs position, combinations between colours and poses were considered as a further research direction. Furthermore, HAAD problems were still not considered.

In Chapter 5, the conclusive studies regarding pose-driven HAR and HAAD were conducted. The first goal was to define and study novel models for joint RGB-poses based HAR. This study provided evidence that multimodal data based on human colours and poses are generally beneficial compared to the unimodal data for HAR. Therefore, the most performing model, among those proposed for HAR, was selected to conduct further studies on HAAD. Thus, the joint RGB-pose based Model B was considered as a multimodal feature extraction method to simultaneously perform HAR and semi-supervised HAAD. The experimentation was conducted on newly recorded multi-target datasets, to compensate for the lack of datasets in the literature regarding posture-related HAAD problems. This experimentation also provided evidence that ActionXPose, as well as its multimodal extension, i.e. Model B, are effective in non-cooperative scenarios for HAR and HAAD. Furthermore, the HAAD considered problem included objects position-related anomalies alongside human posture-related anomalies. The proposed model outperformed state-of-the-art, autoencoder-based baselines on tested datasets. Overall, the proposed model can perform combined joint RGB-pose HAR and, simultaneously, pose-driven and object-positions based HAAD in multi-target, non-cooperative scenarios. To the best of the know-

ledge, this algorithm constitutes a unique example in the literature with such characteristics. The effectiveness of the proposed model is also showed in this video examples[1,2].

Overall, Model B for HAR, proposed in Section 5.2, deployed as part of the bilevel HAAD model proposed in Section 5.3, thoroughly achieve the goals of this thesis.

6.1 Future Work

This thesis's work spanned three years. During this time, OpenPose was not the only main breakthrough advance in the state-of-the-art. Other subsequent and very recent advances suggest exciting research directions which might lead to this thesis further development.

For example, it is evident that OpenPose has limitations, particularly in in-the-wild scenarios. OpenPose's authors are actively developing their algorithm to make it progressively more effective and robust. However, it could be potentially more effective to consider the recent algorithms for image-segmentation, in place or in collaboration with OpenPose. As a matter of fact, image-segmentation detectors can provide human-silhouettes for posture-related HAR and HAAD. Therefore, modern approaches for silhouette-based HAR and HAAD can be studied based on image-segmentation data. Thus, the approaches and the conclusions drawn in Chapter 3 will be necessarily reconsidered.

Another example of future work direction could be extending ActionXPose results on 3D human-skeleton HAR and HAAD based on stereo cameras. In fact, new and very recent stereo cameras can interpolate depth information from a pair of 2D coloured images, providing the third dimension to human poses. Thus, straightforward 3D extensions of ActionXPose

[1] https://www.youtube.com/watch?v=7_mcWCB76Ps
[2] https://www.youtube.com/watch?v=VJD9tYPzHPQ

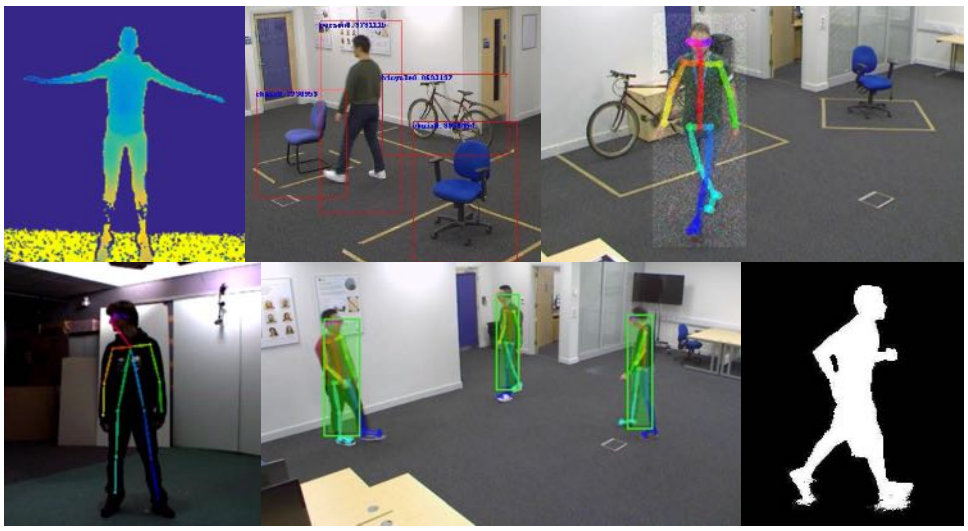


Figure 6.1: This thesis focuses on multi-semantic RGB-based data such as target bounding boxes, object bounding boxes and labels, human silhouettes and human poses.

are suggested.

Future work might also concentrate on testing the proposed approaches on a variety of additional contexts for healthcare and surveillance applications. Furthermore, this thesis raises attention over the current trend in deep learning research, i.e. multi-tasking network training strategies and fusion mechanisms. Therefore, future work might be upon exploring novel fusion mechanism for multi-tasking learning.

It is well known that *multimodality* is, in principle, advantageous for HAR and HAAD. The idea behind the multimodality is that different sensors, e.g. RGB, depth sensors, and accelerometers, can potentially be deployed to compensate to each other limitations, providing robust and effective information. However, multimodal data can be difficult to be obtained due to the non-invasive constraints.

On the other hand, as suggested by Hampapur et al. [147] and Porter et al. [148], multimodality can also be in terms of multi-scale or multi-semantic data. Multi-semantic processing is performed by humans subconsciously while solving complex tasks such as human action recognition or behaviour

analysis. For example, a human operator observing a hall entrance, not only controls where people are positioned in time and space, i.e. human *tracking*, but also what they are doing, what is their appearance, facial expressions and overall behaviour coherence. This information is retrieved from a single source, i.e. visual data. Therefore, the challenge is *how to mimic such multimodal human ability?* The problem is further exacerbated by the so-called *semantic gap*. While the data acquisition is a very physical phenomenon, the object of interest is defined in very abstract terms, and it is related to a subset of the full carried information (what the human operator considers interesting). Therefore, we would need to integrate different methods considering them as components of a more general hierarchy model [148]. Following this purpose, this thesis exploited RGB data to extract multi-semantic information, e.g. *target bounding boxes*, *object bounding boxes and labels*, *human silhouettes* and *human poses* (Figure 6.1). However, multimodal generalisation of the models proposed in this thesis for multimodal implementations can be further explored. Last but not least, since RGB cameras were the main in-

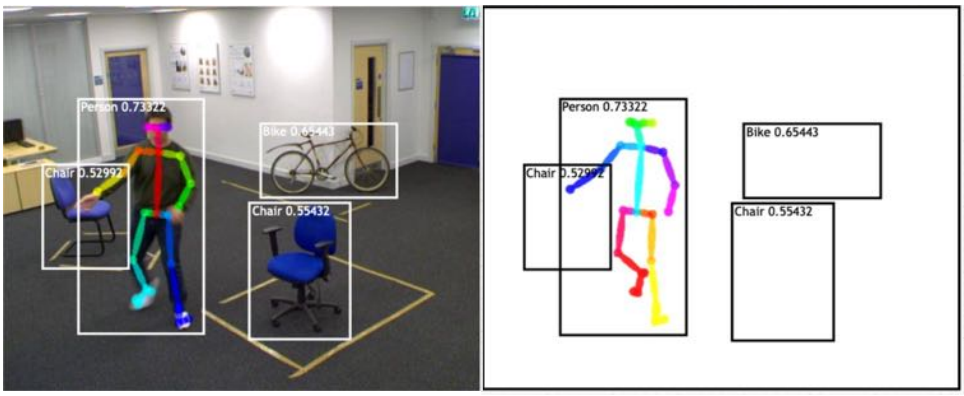


Figure 6.2: Example of privacy-preserving data. Human poses and bounding boxes are the only stored and processed data, while RGB data are discarded and neglected from the processing.

put source in this thesis, privacy issues need to be considered. In fact, RGB data contains sensitive information, such as target's identity, that might compromise the applicability of the proposed methods. Consequently, the

above-mentioned multi-semantics data acquisition must take into account that the target's privacy might be not questioned, depending on the desired application. Thus, a trade-off between information and data protection is required. In other words, it is desirable to find RGB-based solutions which are still effective even if the privacy of the target is preserved. In this thesis, preliminary approaches for privacy-preserving processing has been explored by considering human poses and discarding RGB data right after target detection (Figure 6.2). However, still target identification can be, in principle, performed by using human poses. Therefore, more sophisticated solutions for privacy-preserving HAR and HAAD can be further studied.

References

[1] "GizmoSupport," *https://gizmosupport.com/blog/importance-of-cctv-surveillance-monitoring/*, 2020.

[2] "Wikipedia," *https://en.wikipedia.org/wiki/Kerch_Polytechnic_College_-massacre*, 2020.

[3] "Digital Trends," *https://www.digitaltrends.com/home/google-home-vs-amazon-echo/*, 2020.

[4] "Blendspace," *https://www.tes.com/lessons/fnfPWa-MBt24lQ/brett-stebelton-fear-project*, 2020.

[5] B. Langmann, K. Hartmann, and O. Loffeld, "Depth Camera Technology Comparison and Performance Evaluation," *International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, vol. 2, pp. 438–444, 2012.

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

[7] C. Schuldt, L. Barbara, and S. Stockholm, "Recognizing Human Actions: A Local SVM Approach," *International Conference on Pattern Recognition (ICPR)*, vol. 3, pp. 32–36, 2004.

[8] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, "The i3DPost

multi-view and 3D human action/interaction database," in *European Conference for Visual Media Production (CVMP)*, pp. 159–168, 2009.

[9] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, pp. 249–257, 2006.

[10] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single Shot MultiBox Detector," in *ECCV*, 2016.

[11] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv*, apr 2018.

[12] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7291–7299, 2017.

[13] I. Goodfellow, *Deep learning*. The MIT Press, 2016.

[14] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PLoS ONE*, vol. 11, no. 4, 2016.

[15] I. T. J. Springer, "Principal Component Analysis, Second Edition," tech. rep.

[16] E. Hyvarinen, Aapo Karhunen, Juha Oja, *Independent Component Analysis*. Hoboken, Wiley, 2004.

[17] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "Training algorithm for optimal margin classifiers," in *ACM Workshop on Computational Learning Theory*, (New York, New York, USA), pp. 144–152, Publ by ACM, 1992.

[18] C. Campbell and Y. Ying, "Learning with support vector machines," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 10, pp. 1–95, 2011.

[19] J. Carreira and A. Zisserman, "Quo Vadis, action recognition? A new model and the kinetics dataset," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, pp. 4724–4733, 2017.

[20] B. K. Lavine, "Clustering and classification of analytical data," in *Encyclopedia of Analytical Chemistry*, pp. 1–21, 2000.

[21] T. Kohonen, *Self-organizing maps*. Berlin: Springer, 3rd ed., 2001.

[22] J. Vesanto and E. Alhoniemi, "Clustering of the Self-Organizing Map," Tech. Rep. 3, 2000.

[23] 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, pp. 1–31, 2017.

[24] E. J. Fernandez-Sanchez, J. Diaz, and E. Ros, "Background subtraction based on color and depth using active sensors," *Sensors*, vol. 13, no. 7, pp. 8895–8915, 2013.

[25] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image and Vision Computing*, vol. 60, pp. 4–21, 2017.

[26] D. Weinland, M. Özuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," in *LNCS - Lecture Notes in Computer Science*, vol. 6313, pp. 635–648, 2010.

[27] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "RGB-D-based human motion recognition with deep learning: A survey," *Computer Vision and Image Understanding*, vol. 171, pp. 118–139, 2018.

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

[29] J. Schmidhuber, "Deep Learning in Neural Networks: An Overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[30] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A Short Note on the Kinetics-700 Human Action Dataset," *arXiv*, 2019.

[31] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild," *arXiv*, 2012.

[32] G. Singh, S. Saha, M. Sapienza, P. Torr, and F. Cuzzolin, "Online Real-Time Multiple Spatiotemporal Action Localisation and Prediction," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 3657–3666, 2017.

[33] M. Cristani, A. Pesarin, A. Vinciarelli, M. Crocco, and V. Murino, "Look at Who's Talking: Voice Activity Detection by Automated Gesture Analysis," *Proceedings of Interhub*, vol. 227, pp. 72–80, 2011.

[34] M. Cristani, R. Raghavendra, A. Del Bue, and V. Murino, "Human behavior analysis in video surveillance: A Social Signal Processing perspective," *Neurocomputing*, vol. 100, pp. 86–97, 2013.

[35] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[36] L. Bazzani, M. Cristani, D. Tosato, M. Farenzena, G. Paggetti, G. Menegaz, and V. Murino, "Social interactions by visual focus of attention in a three-dimensional environment," in *Expert Systems*, vol. 30, pp. 115–127, 2013.

[37] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino, "Social interaction discovery by statist-

ical analysis of F-formations," *British Machine Vision Conference (BMVC)*, pp. 23.1–23.12, 2011.

[38] H. Hung and B. Kröse, "Detecting F-formations as dominant sets," *International Conference on Multimodal Interfaces*, pp. 231–238, 2011.

[39] M. Van Droogenbroeck and O. Barnich, "ViBe: A Disruptive Method for Background Subtraction," in *Background Modeling and Foreground Detection for Video Surveillance*, no. July, pp. 7.1–7.23, 2014.

[40] G. Batchuluun, Y. Kim, J. Kim, H. Hong, and K. Park, "Robust Behavior Recognition in Intelligent Surveillance Environments," *Sensors*, vol. 16, no. 7, p. 1010, 2016.

[41] N. C. Tang, Y.-Y. Lin, J.-H. Hua, M.-F. Weng, and H.-Y. M. Liao, "Human action recognition using associated depth and skeleton information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4608–4612, 2014.

[42] N. C. Tang, Yen-Yu Lin, Ju-Hsuan Hua, Shih-En Wei, Ming-Fang Weng, and H.-Y. M. Liao, "Robust Action Recognition via Borrowing Information Across Video Modalities," *IEEE Transactions on Image Processing*, vol. 24, no. 2, pp. 709–723, 2015.

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

[44] D. Weinland and E. Boyer, "Action recognition using exemplar-based embedding," in *26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 1–7, IEEE, 2008.

[45] M. B. M. Holte, T. B. Moeslund, and C. Tran, "Human action recognition using multiple views: a comparative perspective on recent developments,"

ACM Workshop on Human Gesture and Behavior Understanding, pp. 47–52, 2011.

[46] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. I, pp. 886–893, 2005.

[47] A. Klaser, M. Marszalek, and C. Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," *British Machine Conference*, pp. 99.1–99.10, 2008.

[48] C. Li, Y. Liu, J. Wang, and H. Wang, "Combining Localized Oriented Rectangles and Motion History Image for Human Action Recognition," in *International Symposium on Computational Intelligence and Design (ISCID)*, pp. 53–56, 2014.

[49] M. E. Kundegorski and T. P. Breckon, "Posture estimation for improved photogrammetric localization of pedestrians in monocular infrared imagery," *Optics and Photonics for Counterterrorism, Crime Fighting, and Defence XI; and Optical Materials and Biomaterials in Security and Defence Systems Technology XII*, vol. 9652, 2015.

[50] F. Liu, X. Xu, S. Qiu, and C. Qing, "Simple to Complex Transfer Learning for Action Recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 2, pp. 949–960, 2016.

[51] 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.

[52] J. K. Aggarwal, "Human Activity Analysis: A Review," *ACM Computing Surveys*, vol. 43, no. 3, pp. 1–43, 2011.

[53] M. Scherer, M. Walter, and T. Schreck, "Histograms of oriented gradients for 3d object retrieval," *Proceedings of the WSCG*, 2010.

[54] R. Dupre, V. Argyriou, D. Greenhill, and G. Tzimiropoulos, "A 3D Scene Analysis Framework and Descriptors for Risk Evaluation," in *IEEE International Conference on 3D Vision*, pp. 100–108, 2015.

[55] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Computer Vision and Pattern Recognition (CVPR)*, pp. 716–723, 2013.

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

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

[58] S. Azary and A. Savakis, "Multi-view action classification using sparse representations on Motion History Images," in *Western New York Image Processing Workshop (WNYISPW)*, pp. 5–8, 2012.

[59] 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. 15283–15296, 2018.

[60] A. A. Chaaraoui and F. Flórez-Revuelta, "Optimizing human action recognition based on a cooperative coevolutionary algorithm," *Engineering Applications of Artificial Intelligence*, vol. 31, pp. 116–125, may 2014.

[61] C.-b. Jin, S. Li, and H. Kim, "Real-Time Action Detection in Video Surveillance using Sub-Action Descriptor with Multi-CNN," *arXiv*, 2017.

[62] B. Liang and L. Zheng, "A Survey on Human Action Recognition Using

Depth Sensors," in *International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–8, 2015.

[63] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang, "RGB-D-based action recognition datasets: A survey," *Pattern Recognition*, vol. 60, pp. 86–105, 2016.

[64] P. Wang and P. O. Ogunbona, "RGB-D-based Motion Recognition with Deep Learning: A Survey," *International Journal of Computer Vision (IJCV)*, no. June, pp. 1–34, 2017.

[65] E. Insafutdinov, L. Pishchulin, B. Andres, M. Andriluka, and B. Schiele, "Deepercut: A deeper, stronger, and faster multi-person pose estimation model," in *Lecture Notes in Computer Science*, vol. 9910 LNCS, pp. 34–50, 2016.

[66] "Medium," *https://medium.com/@erica.z.zheng/installing-openpose-on-ubuntu-18-04-cuda-10-ebb371cf3442*, 2020.

[67] Y. Guo, Y. Li, and Z. Shao, "RRV: A Spatiotemporal Descriptor for Rigid Body Motion Recognition," *IEEE Transactions on Cybernetics*, vol. 48, pp. 1513–1525, may 2018.

[68] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1290–1297, 2012.

[69] A. Taha, H. H. Zayed, and M. E. Khalifa, "Skeleton-based Human Activity Recognition for Video Surveillance," *International Journal of Scientific & Engineering Research*, vol. 6, no. 1, pp. 993–1004, 2015.

[70] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI Conference on Artificial Intelligence*, pp. 7444–7452, 2018.

[71] J. Liu, A. Shahroudy, D. Xu, A. Kot Chichung, and G. Wang, "Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[72] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 07-12-June, pp. 1110–1118, 2015.

[73] V. Veeriah, Z. Naifan, and G.-J. Qi, "Differential Recurrent Neural Networks for Action Recognition," in *International Conference on Computer Vision (ICCV)*, pp. 4041–4049, 2015.

[74] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence Feature Learning for Skeleton based Action Recognition using Regularized Deep LSTM Networks," in *Conference on Artificial Intelligence (AAAI)*, pp. 3697–3703, 2016.

[75] A. Kovashka and K. Grauman, "Learning a Hierarchy of Discriminative Space-Time Neighborhood Features for Human Action Recognition," *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2046–2053, 2010.

[76] Y. Zhang, X. Liu, M. C. Chang, W. Ge, and T. Chen, "Spatio-temporal phrases for activity recognition," in *European Conference on Computer Vision (ECCV)*, pp. 707–721, 2012.

[77] S. Ji, M. Yang, K. Yu, and W. Xu, "3D convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–31, 2013.

[78] R. Almeida, Z. K. Goncalves Do Patrocinio, and S. J. F. Guimaraes, "Exploring quantization error to improve human action classification," in *In-*

ternational Joint Conference on Neural Networks (IJCNN), pp. 1354–1360, 2017.

[79] M. Vrigkas, V. Karavasilis, C. Nikou, and I. A. Kakadiaris, "Matching mixtures of curves for human action recognition," *Computer Vision and Image Understanding*, vol. 119, pp. 27–40, 2014.

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

[81] M. Raptis and S. Soatto, "Tracklet descriptors for action modeling and video analysis," in *European Conference on Computer Vision (ECCV)*, pp. 577–590, 2010.

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

[83] A. Gilbert, J. Illingworth, and R. Bowden, "Action Recognition Using Mined Hierarchical Compound Features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 883–897, 2011.

[84] Z. Fu, F. Angelini, J. Chambers, and S. M. Naqvi, "Multi-Level Cooperative Fusion of GM-PHD Filters for Online Multiple Human Tracking," *IEEE Transactions on Multimedia*, vol. 21, no. 9, pp. 2277–2291, 2019.

[85] Z. Fu, P. Feng, F. Angelini, J. Chambers, and S. M. Naqvi, "Particle PHD Filter Based Multiple Human Tracking Using Online Group-Structured Dictionary Learning," *IEEE Access*, vol. 6, pp. 14764–14778, 2018.

[86] A. Ur-Rehman, S. M. Naqvi, L. Mihaylova, and J. A. Chambers, "Multi-Target Tracking and Occlusion Handling With Learned Variational Bayesian

Clusters and a Social Force Model," *IEEE Transactions on Signal Processing*, vol. 64, no. 5, pp. 1320–1335, 2016.

[87] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 3464–3468, 2016.

[88] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[89] S. O. Ba and J. M. Odobez, "A study on visual focus of attention recognition from head pose in a meeting room," in *Lecture Notes in Computer Science*, vol. 4299 LNCS, pp. 75–87, Springer Berlin Heidelberg, 2006.

[90] R. Stiefelhagen, M. Finke, J. Yang, and A. Waibel, "From Gaze to Focus of Attention," *Proceedings of Workshop on Perceptual User Interfaces*, vol. 1614, pp. 25–30, 1999.

[91] K. Smith, S. Ba, J.-M. Odobez, and D. Gatica-Perez, "Tracking the Visual Focus of Attention for a Varying Number of Wandering People," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1212–1229, 2008.

[92] Y. Matsumoto, T. Ogasawara, and A. Zelinsky, "Behavior recognition based on head pose and gaze direction measurement," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 3, pp. 2127–2132, IEEE, 2002.

[93] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing based on multiple cues," *IEEE Transactions on Neural Networks*, vol. 13, no. 4, pp. 928–938, 2002.

[94] N. M. Robertson and I. D. Reid, "Automatic Reasoning about Causal Events in Surveillance Video," *EURASIP Journal on Image and Video Processing*, vol. 2011, no. 1, pp. 1–19, 2011.

[95] C. Chen, A. Heili, and J. M. Odobez, "A joint estimation of head and body orientation cues in surveillance video," in *International Conference on Computer Vision (ICCV)*, pp. 860–867, IEEE, 2011.

[96] F. Flohr, M. Dumitru-Guzu, J. F. P. Kooij, and D. M. Gavrila, "Joint probabilistic pedestrian head and body orientation estimation," *IEEE Intelligent Vehicles Symposium*, vol. 16, no. 4, pp. 1872 – 1882, 2015.

[97] E. Ricci, J. Varadarajan, R. Subramanian, S. R. Bulo, N. Ahuja, and O. Lanz, "Uncovering interactions and interactors: Joint estimation of head, body orientation and f-formations from surveillance videos," in *IEEE International Conference on Computer Vision (ICCV)*, vol. 11, pp. 4660–4668, 2016.

[98] "Medgadget," *https://www.medgadget.com/2018/07/humanoid-robot-teaches-autistic-kids-to-recognize-emotions.html*, 2018.

[99] "Github.com," *https://github.com/DhruvJawalkar/PyTorch-OpenPose-Realtime-Multi-Person-2D-Pose-Estimation-using-Part-Affinity-Fields*, 2019.

[100] N. Vaswani, A. Roy-chowdhury, and R. Chellappa, "Shape Activity: A Continuous-State HMM for Moving / Deforming Shapes With Application to Abnormal Activity Detection," *IEEE Transaction on Image Processing*, vol. 14, no. 10, pp. 1603–1616, 2005.

[101] D. G. Kendall, *Shape and shape theory*. Wiley, 1999.

[102] D. G. Kendall, "A Survey of the Statistical Theory of Shape," *Statistical Science*, vol. 4, no. 2, pp. 87–99, 1989.

[103] A. Kale, A. Sundaresan, A. N. Rajagopalan, N. P. Cuntoor, A. K. Roy-Chowdhury, V. Kruger, and R. Chellappa, "Identification of humans using gait," *IEEE Transactions on Image Processing*, vol. 13, no. 9, pp. 1163–1173, 2004.

[104] N. Vaswani, A. Roy Chowdhury, and R. Chellappa, "Activity recognition using the dynamics of the configuration of interacting objects," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. II–633–40, 2003.

[105] N. Vaswani, A. Roy Chowdhury, and R. Chellappa, "Statistical shape theory for activity modeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, pp. 493–496, IEEE, 2003.

[106] R. Chellappa, N. Vaswani, and A. K. R. Chowdhury, "Activity Modeling and Recognition Using Shape Theory," *Behavior Representation in Modeling and Simulation*, pp. 1–3, 2003.

[107] S. Das and N. Vaswani, "Nonstationary shape activities: Dynamic models for landmark shape change and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 579–592, 2010.

[108] W. Sultani, C. Chen, and M. Shah, "Real-World Anomaly Detection in Surveillance Videos," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6479–6488, 2018.

[109] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning Temporal Regularity in Video Sequences," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–31, 2016.

[110] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos us-

ing spatiotemporal autoencoder," in *Lecture Notes in Computer Science*, vol. 10262 LNCS, pp. 189–196, Springer, 2017.

[111] S. Amraee, A. Vafaei, K. Jamshidi, and P. Adibi, "Anomaly detection and localization in crowded scenes using connected component analysis," *Multimedia Tools and Applications*, vol. 77, no. 12, pp. 14767–14782, 2018.

[112] Y. Gao, H. Liu, X. Sun, C. Wang, and Y. Liu, "Violence detection using Oriented VIolent Flows," *Image and Vision Computing*, vol. 48, pp. 37–41, 2016.

[113] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-Cascade: Cascading 3D Deep Neural Networks for Fast Anomaly Detection and Localization in Crowded Scenes," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1992–2004, 2017.

[114] W. Li, V. Mahadevan, and N. Vasconcelos, "Anomaly detection and localization in crowded scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 1, pp. 18–32, 2014.

[115] S. Mohammadi, A. Perina, H. Kiani, and V. Murino, "Angry Crowds: Detecting Violent Events in Videos," in *European Conference on Computer Vision (ECCV)*, 2016.

[116] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision and Image Understanding*, vol. 117, no. 6, pp. 633–659, 2013.

[117] 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)*, pp. 2556–2563, 2011.

[118] "Web Page," *http://mha.cs.umn.edu/*.

[119] C. Lu, J. Shi, and J. Jia, "Abnormal Event Detection at 150 FPS in MATLAB," in *IEEE International Conference on Computer Vision (ICCV)*, pp. 2720–2727, 2013.

[120] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 555–560, mar 2008.

[121] S. A. Velastin and D. A. Gómez-Lira, "People Detection and Pose Classification Inside a Moving Train Using Computer Vision," in *International Visual Informatics Conference*, pp. 319–330, Springer, 2017.

[122] D. G. Lowe, "Distinctive image features from scale invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[123] O. Barnich and M. V. Droogenbroeck, "ViBe : A universal background subtraction algorithm for video sequences," *IEEE Transactions on Image Processing*, vol. 20, no. 6, pp. 1709–1724, 2011.

[124] S. M. M. Ahsan, J. K. Tan, H. Kim, and S. Ishikawa, "Histogram of spatio temporal local binary patterns for human action recognition," in *International Conference on Soft Computing and Intelligent Systems (SCIS) and International Symposium on Advanced Intelligent Systems (ISIS)*, pp. 1007–1011, 2014.

[125] B. Hilsenbeck, D. Munch, H. Kieritz, W. Hubner, and M. Arens, "Hierarchical Hough forests for view-independent action recognition," in *International Conference on Pattern Recognition (ICPR)*, pp. 1911–1916, 2016.

[126] G. Castro-Munoz and J. Martinez-Carballido, "Real Time Human Action Recognition Using Full and Ultra High Definition Video," in *International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 509–514, 2015.

[127] P. Feng, W. Wang, S. M. Naqvi, and J. Chambers, "Adaptive Retrodiction Particle PHD Filter for Multiple Human Tracking," *IEEE Signal Processing Letters*, vol. 23, no. 11, pp. 1592–1596, 2016.

[128] P. Feng, W. Wang, S. Dlay, S. M. Naqvi, and J. Chambers, "Social Force Model-Based MCMC-OCSVM Particle PHD Filter for Multiple Human Tracking," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 725–739, 2017.

[129] F. Karim, S. Majumdar, H. Darabi, and S. Harford, "Multivariate LSTM-FCNs for Time Series Classification," *arXiv*, 2018.

[130] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[131] H. Jo, Y. H. Na, and J. B. Song, "Data augmentation using synthesized images for object detection," in *International Conference on Control, Automation and Robotics (ICCAR)*, pp. 1035–1038, 2017.

[132] W.-N. Hsu, Y. Zhang, and J. Glass, "Unsupervised Domain Adaptation for Robust Speech Recognition via Variational Autoencoder-Based Data Augmentation," *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 16–23, 2017.

[133] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, "MOT16: A Benchmark for Multi-Object Tracking," *arXiv*, 2016.

[134] T.-Y. Lin, C. L. Zitnick, and P. Doll, "Microsoft COCO: Common Objects in Context," *arXiv*, 2015.

[135] L. Fei-Fei, R. Fergus, and P. Perona, "Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories," in *Computer Vision and Pattern Recognition Workshop (CVPR)*, pp. 178–178, 2004.

[136] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *International Conference on Computer Vision (ICCV)*, vol. II, pp. 1458–1465, 2005.

[137] A. Berg, T. Berg, and J. Malik, "Shape Matching and Object Recognition Using Low Distortion Correspondences," in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, pp. 26–33, 2005.

[138] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer-Verlag New York, 2013.

[139] D. Roussinov and H. Chen, "A scalable self-organizing map algorithm for textual classification: A neural network approach to thesaurus generation," *Communication Cognition and Artificial Intelligence (CC-AI)*, vol. 15, no. 1-2, pp. 81–111, 1998.

[140] K. Simonyan and A. Zisserman, "Two-stream Convolutional Networks for Action Recognition in Videos," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, pp. 568–576, 2014.

[141] R. Zhao, H. Ali, and P. van der Smagt, "Two-stream RNN/CNN for action recognition in 3D videos," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4260–4267, 2017.

[142] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv*, 2015.

[143] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, pp. 211–252, dec 2015.

[144] C. Campbell, *Learning with support vector machines*. Morgan & Claypool, 2011.

[145] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *IEEE International Conference on Computer Vision*, vol. 2017-Octob, pp. 618–626, 2017.

[146] E. B. Nievas, O. D. Suarez, G. B. Garcia, and R. Sukthankar, "Hockey Fight Detection Dataset," in *Computer Analysis of Images and Patterns*, pp. 332–339, 2011.

[147] A. Hampapur, L. Brown, J. Connell, A. Ekin, N. Haas, M. Lu, H. Merkl, and S. Pankanti, "Smart video surveillance: exploring the concept of multiscale spatiotemporal tracking," *IEEE Signal Processing Magazine*, vol. 22, no. 2, pp. 38–51, 2005.

[148] R. Porter, A. M. Fraser, and D. Hush, "Wide-area motion imagery," *IEEE Signal Processing Magazine*, vol. 27, no. 5, pp. 56–65, 2010.