POPULAR DEMOCRATIC REPUBLIC OF ALGERIA
HIGHER EDUCATION AND SCIENTIFIC RESEARCH'S MINISTRY

Faculty of Exact Sciences, Nature Sciences and Life
Department of Mathematics and Computer Science
Larbi Ben M'hidi University, Oum El Bouaghi, Algeria

Vision-based Human Activities Recognition in Supervised or Assisted Environment

Submitted submitdate, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Computer Science

Beddiar Djamila Romaissa

Supervised by:
Mr **Nini Brahim** **Professor**

Jury members:
Mr Bourouis Abdelhabib **Professor** **Larbi Ben M'hidi University, Oum El Bouaghi**
Mr Marir Toufik **MCA** **Abd Elhamid Mahri University, Constantine 2**
Mr Mourad BOUZNADA **MCA** **Université A. Hamid Mehri, Constantine 2**

February 2021

Abstract

Human Activity Recognition HAR has been a hot research topic in the last decade due to its wide range of applications. Indeed, it has been the basis for implementation of many computer vision applications, home security, video surveillance, and human-computer interaction. We intend by HAR, tools, and systems allowing to detect and recognize actions performed by individuals. With the considerable progress made in sensing technologies, HAR systems shifted from wearable and ambient-based to vision-based. This motivated the researchers to propose a large mass of vision-based solutions. From another perspective, HAR plays an important role in the health care sector and gets involved in the construction of fall detection systems and many smart home-related systems. Fall detection FD consists in identifying the occurrence of falls among other daily life activities. This is essential because falling is one of the most frequent serious health issues encountered by seniors. FD systems are especially used in elderly homes and workplaces to enable elderly isolated populations to live alone for as long as possible, enhance their security and remote assistance.

In this thesis, gaps in HAR field and current challenges are identified. This was performed by reviewing the most prominent state-of-the-art techniques, analyzing and evaluating them. Based on the literature review, new algorithms are introduced and embedded to explore the multi-modal HAR by combining different modalities that allowed us to highlight the spatial and temporal evolution of the actions. The proposed approach based on deep learning and video representation is quite simple and achieves state-of-the-art results.

Afterwards, to address some issues related to FD, we combine human body geometry available at different frames of the video sequence with pose estimation. The proposed approach relies on deep learning architectures and transfer learning to achieve high accuracy while identifying falls from daily life activities and is intended to be used for elderly assistance. Finally, the thesis identifies mandatory extensions regarding our proposed frameworks for HAR and FD and future research directions.

Résumé

La reconnaissance des actions humaines est devenue un sujet scientifique en pleine effervescence grace à ses divers domaines d'application. Elle est à la base du développement de nombreuses applications d'interaction homme-machine, vision artificielle, sécurité, vidéosurveillance et assistance à domicile. La reconnaissance des actions humaines est l'ensemble des outils et systèmes permettant de détecter et de reconnaitre l'action réalisée par l'individu. L'évolution remarquable qu'ont connu les technologies de détection ces dernières années a influencé de manière directe le développement des systèmes de reconnaissance des activités humaines. Ceci a permis de passer des systèmes à base de contact aux systèmes à base de vision, ce qui a motivé les chercheurs à proposer une grande masse de solutions. Par ailleurs, la reconnaissance des actions humaines joue un rôle primordial dans le secteur de la santé et l'assistance à domicile. Elle est exploitée dans la construction des systèmes de détection de chutes ainsi que d'autres systèmes relatifs aux maisons intelligentes. La détection des chutes consiste à identifier l'occurrence de chutes parmi les différentes actions de vie quotidienne. Ceci est essentiel car la chute est considérée comme l'un des problèmes de santé auxquels les seniors sont fréquemment exposés. Les systèmes de détection de chutes sont particulièrement utilisés dans les maisons et les bureaux des seniors pour leur permettre de vivre independamment de façon autonome aussi longtemps que possible, optimiser leur sécurité et améliorer les services d'assistance à distance.

A l'issue de notre synthèse de l'état de l'art relatif au domaine de la reconnaissance des actions humaines, il nous a été possible d'identifier les challenges y afférent, d'analyser et d'évaluer les techniques existantes et par conséquent, mettre en avant quelques lacunes de recherche que nous proposons d'étudier dans ce travail. A cet effet, de nouveaux algorithmes sont proposés et sont introduits pour explorer la reconnaissance des actions humaines en combinant différentes modalités de données. Ceci nous a permis également de mettre en évidence la combinaison de l'évolution spatiale et temporelle de l'action. L'approche que nous proposons est basée sur l'apprentissage profond et la représentation de vidéo. Elle est simple et démontre de très bonnes performances.

Par ailleurs, pour résoudre quelques problèmes liés à la détection de chutes, nous combinons la géométrie du corps humain, disponible à travers les différentes séquences vidéo, avec l'estimation de poses. L'approche proposée, fondée sur l'apprentissage profond et le transfert d'apprentissage, permet d'atteindre un haut niveau de précision par une meilleure identification des chutes liées à l'exercice des activités quotidiennes. Elle est ainsi destinée à l'assistance des seniors dans leur vie quotidienne. Enfin, cette thèse identifie d'autres perspectives futures de recherche et des extensions triviales aux approches proposées pour la reconnaissance des actions humaines et des chutes.

ملخص

في العقد الأخير، كان مجال التعرف على النشاط البشري موضوع بحث ساخن بسبب إستعمالاته في مجموعة واسعة من التطبيقات. و بالفعل ، لقد كان الأساس لتنفيذ العديد من تطبيقات الرؤية الحاسوبية، الأمن المنزلي، المراقبة بالفيديو، أو التفاعل بين البشر والكمبيوتر. نقصد من خلال مجال التعرف على النشاط البشري الأدوات أوالأنظمة الملحقة التي تسمح باكتشاف والتعرف على النشاط الذي يقوم به الفرد. مع التقدم الكبير الذي تم إحرازه في تقنيات الإستشعار، تحولت أنظمة التعرف على النشاط البشري من الأجهزة القابلة للإرتداء و المحيطة بالإسناد إلى تلك القائمة على مجال الرؤية. هذا ما حفز الباحثين على إقتراح مجموعة كبيرة من الحلول القائمة على مجال الرؤية. من منظور آخر، يلعب مجال التعرف على النشاط البشري دوراً هاما في قطاع الرعاية الصحية، كما أستعمل أيضا في بناء أنظمة الكشف عن السقوط والعديد من الأنظمة المتعلقة بالمنازل الذكية. يكمن نظام الكشف عن السقوط من تحديد دقيق لحالة السقوط بين باقي أنشطة الحياة اليومية. يعتبر هذا النظام جد ضروري خاصة أن السقوط هو أحد أكثر المشاكل الصحية الخطيرة التي يواجهها كبار السن. تُستخدم أنظمة الكشف عن السقوط بشكل خاص في دور المسنين وأماكن العمل لمساعدة المسنين المعزولين عن بُعد و تعزيز أمنهم وتمكينهم من العيش بمفردهم لأطول فترة ممكنة.

في هذه الأطروحة، تم تحديد الفجوات في مجال التعرف على النشاط البشري في ما يخص جميع التحديات الحالية. وقد تم إجراء ذلك من خلال مراجعة أبرز البحوث العلمية وأيضا جميع التقنيات الحديثة مع تحليلها وتقييمها. و بناءً على ما تحصلنا عليه من مراجعتها، تم تقديم خوارزميات جديدة ودمجها لاستكشاف متعدد الوسائط، مجال التعرف على النشاط البشري وذلك من خلال الجمع بين الطرق المختلفة التي سمحت لنا بتسليط الضوء على التطور المكاني والزماني للأنشطة التي يقوم بها الأشخاص. ومن هنا اقترحنا النهج المتبع في هذا البحث والذي هو قائم على استعمال التعلم العميق وتمثيل بالفيديو. الطريقة المقترحة تعتبر بسيطة للغاية و أيضا توصلنا من خلالها الى تحقيق نتائج جيدة.

بعد ذلك ،و لغرض معالجة بعض المشكلات المتعلقة بأنظمة الكشف عن السقوط ، نقوم بدمج هندسة جسم الإنسان المتوفرة في أوقات مختلفة من تسلسلات الفيديو مع تقدير للوضعية. يعتمد النهج المقترح على معماريات التعلم العميق وأيضا تقنية نقل التعلم وذلك لتحقيق دقة عالية أثناء تحديد السقوط من أنشطة الحياة اليومية الأخرى ويهدف من استخدامه خاصة لمساعدة المسنين. في الأخير، نحدد من خلال البحوث المنجزة في هذه الأطروحة إلى الإمتدادات الإلزامية فيما يتعلق بالأطرالمقترحة سواءاً في مجال التعرف على النشاط البشري، في أنظمة الكشف عن السقوط، أو إتجاهات البحث المستقبلية.

Dedication

To the fond memory of my father, who passed away very early and could not grab the fruit of his success. He had always been my source of inspiration, courage and strength. To the memory of my father in law, who pushed me to finish this thesis. He passed away after the edition of this thesis and couldn't read these lines ! I wish they could still be alive today to share with me the celebration and the success of my graduation with a Doctor of Philosophy degree. May their memory be eternal;

To my mother, who worked hard to raise us all by herself. She is the reason of what I become today. My mother infused me with love, encouragement and prays of day and night which helped me a lot to make this dream a reality;

To my beloved husband Ayyoub, who has drown me in love, patience, kindness and support. I will always appreciate all he has done for me;

To Chafia, my second mother, who also scarified a lot. Her support, continuous care and unconditional love were the key to all my achievements;

To my brothers, Zoheir and Haider and my sisters in law, especially Sabrina. I am really grateful to all of them. They have always been a source of strength, inspiration and affection;

To all my beloved family and my husband's family for their endless love, prayers, supports and advice;

To my friends, (can't cite all of you !!) who never left my side. A special feeling of gratitude belongs to each one of them;

This work is a sign of my love to you, all of you.

Acknowledgements

First and foremost, I would like to thank Allah the Almighty, who is most beneficent and merciful for giving me the strength and courage to complete this thesis.

I would like to express my sincere gratitude to my advisor Prof. Nini Brahim, for his continuous encouragement, availability and support of my Ph.D study. His patience, motivation and guidance allowed me to realize and edit this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

Besides my advisor, I would like to deeply acknowledge Prof. Oussalah Mourad, for his insightful comments and encouragement, but also for the confidence he has placed in me. His attention and help incentivized me to widen my research from various perspectives. He gave me the opportunity to work in his team, allowed me to produce good papers and he made me a researcher.

I would also like to thank Dr. Belkebir Djalila, who helped me a lot in finalizing this thesis and who was always there for me to answer my queries and clarify my doubts.

My sincere thanks also goes to Prof. Hadid Abdenour, who provided me an opportunity to join their team as intern in 2018, and who enlightened me the first glance of research.

A big thanks to the team of the Research Laboratory on Computer Science's Complex Systems, and the administration of Larbi Ben Mhidi university for all the support they have been able to transmit to me during these years, and also to be always there to guide me to find the right path by their wisdom and precious advice.

In particular, I am grateful to the entire team of the CMVS, university of Oulu, Finland, who gave me the opportunity to benefit from the laboratory and its research facilities. Without their precious support it would not be possible to conduct this research.

Finally, I thank the members of the jury for agreeing to evaluate our modest work, as well as all those who contributed directly or indirectly to the editing of this thesis.

Contents

List of Figures

List of Tables

Glossary

AAL	Ambient Assisted Living
ADL	Activities of Daily Living
AlexNet	Pretrained model AlexNet
AUC	Area Under the Curve
BN	Bayesian Network
BOW	Bag of Words
CCA	Canonical Correlation Analysis
CNN	Convolutional Neural Network
CRF	Conditional Random Field Models
DI	Dynamic Image
DL	Deep Learning
DNN	Deep Neural Network
ELM	Extreme Learning Machine
FD	Fall Detection
FPR	False Positive Rate
GAN	Generative Adversarial Network
GP	Genetic Programming
HA	Human Activity
HAR	Human activity recognition
HCI	Human Computer Interaction
HMM	Hidden Markov Models
HOF	Histogram of Optical Flow
HOG	Histogram of Oriented Gradients
IoU	Intersection Over Union
KNN	k-Nearest Neighbors Algorithm
LBP	Local Binary Pattern
LSTM	Long Short Term Memory Network
LTC	Long-term Temporal Convolutions
MBH	Motion Boundary Histogram
MEI	Motion Energy Image
MHI	Motion History Image
MSE	Mean Squared Error
NPV	Negative Predictive Value
OF	Optical Flow
PPV	Positive Prediction Value
RBF	Radial Basis Function
Resnet50	Pretrained model Resnet50
RNN	Recurrent Neural Network
SDTD	Sequential Deep Trajectory Descriptor
SIFT	Scale-Invariant Feature Transform
STIP	Spatio-Temporal Interest Points
SVM	Support Vector Machine
VAE	Variational Autoencoders
VGG	Pretrained model VGG

1

General Introduction

Vision-based Human Activity Recognition (HAR) has drawn an important role in the progress of the field of computer vision and machine learning. Human activity recognition refers often to the task of determining and naming activities using sensory observations (1). More specifically, we intend by a human activity (HA), the movement(s) of one or several parts of the individual's body. HA can be either atomic or composed of many primitive actions performed in some sequential order. HAR aims to automatically understand and classify the action performed in the image or the video and attributes the same label if the same action was performed by a different person under different conditions. In other words, it attempts to automatically analyze and recognize HAs using the acquired information from the various types of sensors (2, 3, 4). HAR systems could also be employed to guide subsequent decision-support systems. To this end, the underlined HAR systems are generally preceded by an activity detection task. This consists of the temporal identification and localization of such activity in the scene in a way to boost the understanding of the ongoing event. Therefore, the activity recognition task can be divided into two classes: classification and detection.

1.1 Background and motivations

Human activity recognition has become a hot scientific topic in computer vision community and has drawn much attention of researchers. It is involved in the development of many important applications such as human computer interaction (HCI) (5), virtual reality (6), security (7), video surveillance, home monitoring (8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18) and robotics. Therefore, the wide range of the activity recognition methods is directly linked to the application domain to which they are implemented (10). For instance, video surveillance systems allow to automatically track the crowd and recognize their activities without requiring any strenuous human monitoring. This may help to increase the security in public places and prevent dangerous situations by triggering alerts whenever needed. Giving the ability

to the computer/robot to understand the human activities may facilitate the communication between the human and the computer/robot and makes the robots much more useful when involved in an ongoing activity. Application of HAR in home monitoring permits to monitor the patient's activities, detect any abnormality, offer them efficient health care and thus improve the quality of their life. However, HAR could be burden due to many unresolved challenges such as occlusion, moving and cluttered backgrounds, different illumination conditions, and viewpoint variations (19). Also, the impact of these challenges may vary depending on different factors including the type of the activity under consideration, the viewpoint and the type of the acquisition device, the body parts involved in performing the action and the nature of data whether it is image or video. Generally, activities are categorized into actions, gestures, behaviors, interactions, group actions and events. The complexity of the activity and therefore, the intensity of challenges increase from atomic actions to events. Similarly, challenges' impacts increase when many parts of the body are involved, which include the study of the relationship between them as well. The same thing happens when different views are considered since synchronisation and more data has to be processed. Moreover, analysing videos is much difficult than scrutinizing images since the temporal aspect of the action is also taken into consideration. The motivation behind the work in this thesis is especially related to the wide range of important applications of HAR in the real-world scenarios. This thesis considers actions, interactions with objects and some behaviors. Therefore, the recognition of remaining types of activities are not covered under the scope of this thesis.

Human activity recognition has been studied significantly in the literature. In some previous works (3, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29), HAR systems have been extensively reviewed and discussed. Nevertheless the rapid development of the technology and emergence of new methodological approaches call for a constant update in the field and, thereby, new reviews in a way to benefit the growing HAR community researchers. In this respect, the current thesis completes and updates the aforementioned existing studies.

From another hand, applications of video surveillance have attracted more researchers to propose new techniques of detection and analysis of human activities. Understanding abnormal human activities in monitored environments tend to be very useful and beneficial, especially for public security and elderly monitoring applications. Abnormal activities recognition should be considered as a separate research axis and proposed methods should deal particularly with this kind of activities and their specific characteristics. For instance, detecting falls may help to enhance the elderly life quality, ensure their privacy and boost their independence. This was a motivating factor behind the contributions of this thesis.

1.2 Objectives of the thesis

Recognition of human activities is a hot research topic that has mainly attracted a lot of researchers' attention nowadays. Especially, with the increase development of sensor technologies and deep learning models. Various techniques have been presented in the literature aiming to detect, understand and classify human activities. However, many challenges are still open issues and have to be resolved efficiently. For instance, recognizing actions by fusing different modalities, understanding abnormal activities and detecting falls for elderly are still very challenging and have to be addressed carefully. In this regard, this thesis tries to resolve the following objectives:

- Aim1. Comprehensive review of the state-of-the-art techniques of vision-based human activity recognition.
- Aim2. Extracting and understanding of the limitations of the state-of-the-art techniques. This helps us to identify the gaps for new contributions.
- Aim3. Comprehensive overview of the state-of-the-art techniques of vision-based abnormal human activity recognition to identify limitations and challenges of recognition of such particular activities. Especially, fall detection which is a very important research topic.
- Aim4. Development of a multi-modal framework for human action recognition by combining RGB, Depth and skeleton data. Fusion of multiple modalities is one of the major challenges for human action recognition.
- Aim5. Development of an efficient method for human activity recognition based on a supervised deep learning and a transfer learning model;
- Aim6. Development of an efficient method for human activity recognition based on a new representation of human action in video sequences;
- Aim7. Development of an efficient method for elderly fall detection based on a supervised approach and a transfer learning model;
- Aim8. Comparison between our proposals and existing works in both HAR and FD, using standard benchmark datasets, to produce better results.

1.3 Contributions of the thesis

The main contributions of our thesis could be organized as follows:

- First, we proposed a survey that discusses the most significant advances reported recently in the literature of HAR covering, both the general aspects of human activities recognition and the specific vision-based HAR systems. The Human activity recognition contribution is presented in Chapter 2 and 3, and has been published (30). Our survey differs from the previous works in different aspects:

    1. One can see from Figure 1.1 a that only 9% of the existing surveys from 2010 to present are devoted to discuss the general framework of HAR while 91% are rather presenting specific taxonomies or specific domains. In addition, Figure 1.1 b shows most commonly discussed HAR subjects and the percentage of surveys covering each subject for the ten past years.
    2. This survey paper presented a deeper analysis of human activities recognition by discussing various applications of HAR in different fields, analyzing the proposed approaches, defining abstraction levels of activities and categorizing human action representation methods.
    3. In an attempt to quantify the in-depth characterization of the HAR survey papers reported in our paper, we manually scrutinize the reported survey papers for the last ten years to distinguish in-depth and comprehensive surveys from domain-specific or light surveys. The results are presented in Figure 1.2. It can be seen from this figure that there are only few surveys similar to our work and, thereby, there is a need for updated new comprehensive review for HAR systems. Strictly speaking, the previous quantification (of in depth-survey versus standard-survey) is rather based on the number of technologies and methodologies, structure of taxonomy, in-depth comparative analysis undertaken by the underlined review paper.
    4. In addition to reviewing existing human activities recognition approaches and common related datasets, we classified them according to the modalities used when acquiring data and to the commonly employed three-stage activity recognition process: detection, tracking and classification.
    5. Finally, we identified challenging issues and discussed useful recommendations to provide useful insights to the HAR system development community.

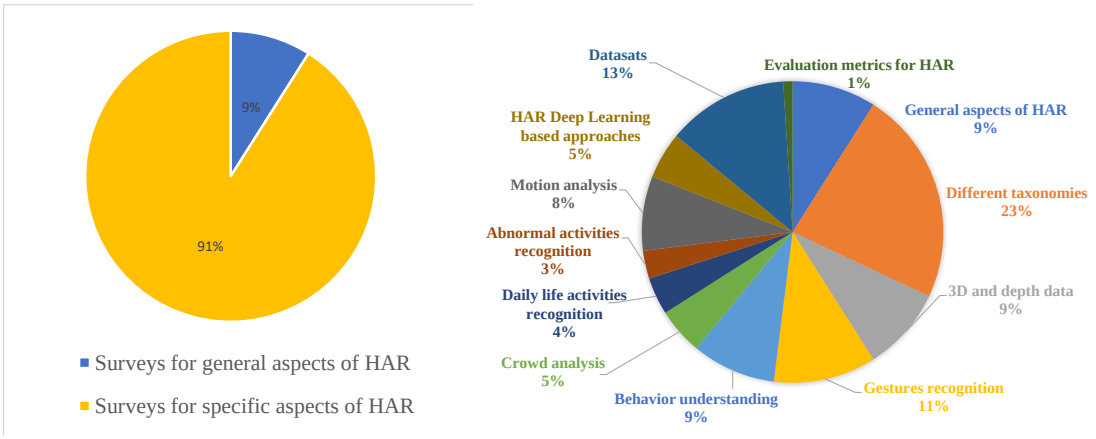


Figure 1.1: HAR related surveys from 2010 until 2019: (a) The percentage of surveys representing general aspects of HAR compared to surveys presenting specific taxonomies and application domains of HAR, (b) Distribution of major subjects of HAR covered by recent surveys

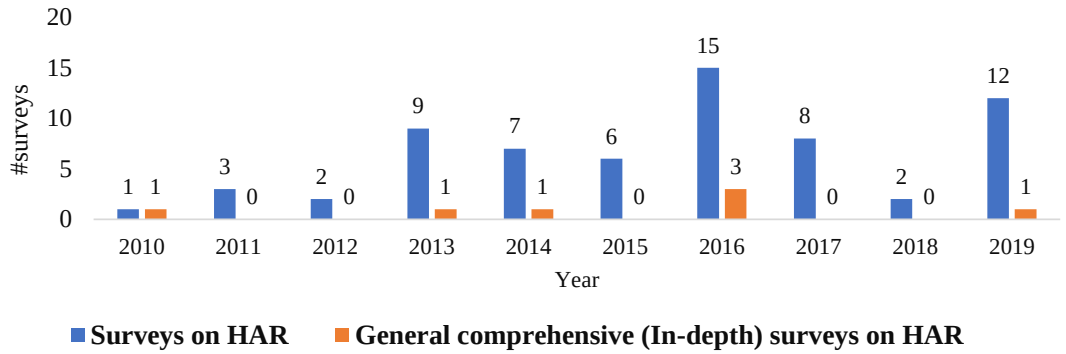


Figure 1.2: Comparison between the number of HAR surveys covered by our study and the number of surveys presenting general and comprehensive analysis of HAR for the period of 2010 to present

- To give a general idea of what exists in the field of abnormal human activity recognition, we provided an overview that includes an analysis of some existing works in this research area. This contribution is presented in Chapters 2,3 and 4, and has been published (4).

- An extension of the above mentioned overview was also proposed by giving update of new research in the field of abnormal human activity recognition focusing on fall detection task. This contribution is presented in Chapter 6, and has been published (31).

- We explored the combination of three modalities (RGB, depth and skeleton data) to design a robust multi-modal framework for vision-based human activity recognition. Illustrative representations of activities using rank pooling were suggested to highlight spatial information, body shape/posture and temporal evolution of actions. Our framework takes advantage of transfer learning from pre-trained models to extract significant features from newly created images. Finally, a Canonical Correlation Analysis inline with a Long Short-Term Memory network are used to fuse extracted features and classify actions from visual descriptive images. This contribution is presented in Chapter 5, and has also been accepted for the 25th International Conference on Pattern Recognition ICPR 2020 (32). Our main contributions can be summarized into the following:

    1. Summarizing RGB and depth videos into dynamic images using an approximate rank pooling method introduced in (33, 34).
    2. Encoding locations of skeleton joints along the video frames into new representations.
    3. Extracting new features from RGB and depth dynamic images and skeleton representations using transfer learning from pre-trained models.
    4. Developing a new feature fusion based strategy using Canonical Correlation Analysis of RGB, depth and skeleton data modalities.
    5. Developing one bi-directional Long Short Term model for action classification that has for input the resulting features fusion vectors.
    6. Evaluating our proposed method on two datasets: UTD-MHAD and NTU RGB+D where we performed cross-view evaluation.

- We presented a fall detection approach based on human body geometry inferred from video sequence frames. The angular information and the distance between the vector formed by the head centroid of the identified facial image and the center hip of the body and the vector aligned with the horizontal axis of the center hip were calculated

and used to construct distinctive image features. Two scenarios were compared. In the first one a two-class SVM was trained on the extracted features from the newly constructed images while in the second scenario a Long Short-Term Memory network (LSTM) was used with the calculated angle and distance sequences to classify falls and no-falls activities. This contribution is presented in Chapter 6, and has also been published within the International Conference on Image Processing Theory, Tools and Applications IPTA 2020 (35). An extension of this work was also published within the International Workshop on Deep Learning for Human-Centric Activity Understanding DL-HAU 2020 (36). It gives a deeper experimentation and comparison aspects. Another extension of this contribution was published by the journal of visual communication and image representation (37). The main contributions of our proposal can be summarized into:

1. A new method for downsampling the videos using optical flow in order to keep only frames with significant motion is put forward. This allows us to reduce the number of frames of the video to be executed in subsequent reasoning.
2. A new research dataset related to our manual annotation task containing the 2D coordinates of both center hip of the body and the head centroid (of facial representation) available from each frame of the video data is made available to research community.
3. A new deep learning-based human pose estimation approach is devised and implemented to automatically annotate the head and the center hip of the body in video sequences.
4. A sound mathematical approach for calculating the angle and the distance between the head centroid and the center hip of the body is demonstrated.
5. The ability to track the variation of the above angular estimation across all frames of the re-sampled video is demonstrated.
6. A new SVM-binary classification that distinguishes fall from non-fall scenarios using the sequence of angles and distances of each video has been put forward.
7. The contribution of other potential feature sets are explored and exploited for representing video sequences.
8. A comparison between LSTM, TCN and SVM classification results has been carried out along with a cross-dataset evaluation.
9. An ablation study was put forward to analyse the importance of each component of the system separately. For that, we focused on feature ablation and hyper-parameters optimization.

1.4 Summary of related publications

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

Journal Papers

- Beddiar Djamila Romaissa, Nini Brahim, Sabokrou Mohammad, Hadid Abdenour. Vision-based human activity recognition: a survey. Multimedia Tools and Applications, 2020, vol. 79, no 41, p. 30509-30555. (30)

- Beddiar Djamila Romaissa, Oussalah Mourad, Nini Brahim. Fall Detection Using Body Geometry and Human Pose Estimation in Video Sequences. Journal of Visual Communication and Image Representation, 2021, p. 103407. (37)

Conference Papers

- Beddiar Djamila Romaissa, Oussalah Mourad, Nini Brahim. Vision-Based Multi-Modal Framework for Action Recognition. In : 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021. p. 5859-5866. (32)

- Beddiar Djamila Romaissa, Oussalah Mourad, Nini Brahim, Bounab Yazid. Fall Detection using Body Geometry in Video Sequences. 2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA). IEEE, 2020. p. 1-5. (35)

- Beddiar Djamila Romaissa, Nini Brahim. Abnormal human activities recognition: brief synthesis of vision based fall detection. 2nd International Conference on Artificial Intelligence and Information Technology ICA2IT 2019. (31)

- Beddiar Djamila Romaissa, Nini Brahim. Vision based abnormal human activities recognition: an overview. 8th International Conference on Information Technology (ICIT) 2017. p.548-553. (4)

Workshop Papers

- Beddiar Djamila Romaissa, Oussalah Mourad, Nini Brahim, Bounab Yazid. Vision-based Fall Detection using Body Geometry. In : International Conference on Pattern Recognition. Springer, Cham, 2021. p. 170-185. (36)

1.5 Organization of the thesis

This thesis consists of six chapters including the current introductory chapter. Chapters 2, 3, 4, 5, and 6 highlight our contributions in human activity recognition and FD fields and are already published in international indexed journals and conferences. A part of the work presented in chapter 6 is currently under review for journal publication. Chapter 7 is devoted to a summary of our study, where we conclude the thesis, address each of the fixed objectives and describe future research directions. Therefore, a brief overview of each chapter is given as follows:

Chapter 2 - Human Activity Recognition: This second chapter provides a global overview of our research area. It clarifies definitions of main concepts related to human activity recognition and highlights general ideas, from different perspectives, of this domain. Moreover, a taxonomy of activity types, the body parts which can be used to identify the various human activities, the nature of the recorded data and the viewpoint of the acquisition device are summarized inline with main human activity recognition application fields. Afterwards, we categorize HAR systems into contact-based and vision-based, where we discuss also drawbacks and advantages of both categories. Before introducing the main area of our contribution, we provide a brief description of the different validation means, the standard and popular benchmarks and evaluation metrics for HAR task.

Chapter 3 - Literature review: Vision-based Human Activity Recognition: In this third chapter, we explored the state-of-the-art of vision-based HAR methods. We provide an overview of survey papers that introduce the recent advances in human activity recognition topic. Subsequently, we propose four taxonomies for classification of HAR approaches according to; (1) the feature extraction process with a categorization of human activity representation methods, (2) the three-stages recognition system followed by the methods used in implementing each of them, (3) the input data modalities and (4) the supervision level of the machine learning. Finally, we discuss some of the existing techniques, enumerate their limitations and the the current challenges they have to cope with to enhance HAR system performances.

Chapter 4 - Overview of Fall Detection: This chapter is devoted to give a general overview of the fall detection field. It clarifies definitions of main concepts related to fall detection, provides a categorization of fall types inline with existing methods in the literature. Applications of FD systems, mainly for elderly monitoring are also highlighted. Afterwards,

we enumerate some FD related benchmark datasets that could be used to evaluate the performance of proposed methodologies. Finally, we discuss limitations and challenges related to FD systems and conclude the chapter.

Chapter 5 - Multi-Modal Human Activity Recognition using deep learning: In this chapter, we explore the combination of three modalities (RGB, depth and skeleton data) to design a robust multi-modal framework for vision-based human activity recognition. Especially, spatial information, body shape/posture and temporal evolution of actions are highlighted using illustrative representations obtained from a combination of dynamic RGB images, dynamic depth images and skeleton data representations. Therefore, each video is represented with three images that summarize the ongoing action. Our framework takes advantage of transfer learning from pre-trained models to extract significant features from these newly created images. Next, we fuse extracted features using Canonical Correlation Analysis and train a Long Short-Term Memory network to classify actions from visual descriptive images. For that, we first review related works in multi-modal vision-based HAR approaches where we also discuss video representations and rank pooling of videos. Then, we detail our proposed and discuss our evaluation findings on the public UTD-MHAD and NTU RGB+D datasets.

Chapter 6 - Vision-based Fall detection using body geometry: We present in this chapter, a fall detection approach that explores human body geometry available at different frames of the video sequence. Especially, the angular information and the distance between the vector formed by the head centroid of the identified facial image, the center hip of the body, and the vector aligned with the horizontal axis of the center hip, are then used to construct distinctive image features. A two-class SVM classifier is trained on the newly constructed feature images, while a Long Short-Term Memory (LSTM) network is trained on the calculated angle and distance sequences to classify falls and non-falls activities. For that, we briefly provide background and previous research related to vision-based Fall detection (FD). Then, we outline our approach and discuss the experimental results of our proposal on the publicly available UR FD and Le2i datasets while comparing a TCN model to the LSTM and SVM classifiers. Finally, an ablation study is put forward to investigate the impact of features and models' parameters on the training process.

Chapter 7 - Summary: We present in this chapter the conclusive summary, the realization of our fixed objectives, by reminding our key contributions, our research and validation methodology. Besides, limitations and future works are also discussed at the end of this chapter.

2

Human Activity Recognition

2.1 Introduction

Human activity recognition (HAR) consists of tools and approaches that allow identifying actions or activities performed by a person or a group of people. It plays a very important role in many real-world applications among which computer vision, human-computer interaction and video surveillance are predominant. Due to the extensive advancement in sensor and visual technology, different approaches of HAR have been implemented, allowing to recognize the human activities in various manners. The authors of (19) categorize HAR systems into visual sensor-based, non visual-sensor based and multi-modal techniques. The first category includes techniques based on visual sensors such as cameras, and depth cameras which record videos and images. Non-visual sensor based techniques rely on the use of sensors that capture data different than images or videos such as inertial data. Finally, the multi-modal techniques combine data recorded using multiple sensors of different modalities. In general. the major difference between these approaches consists in the nature of the recorded data, which may be 2D, 3D images and inertial information. Moreover, to properly understand and recognize the activity, it is interesting to determine the level of the activity as well as the involved part of the body that performs it. This helps to select the better method for recognition of such type of activities. Dealing with data captured from one single view or multiple views or data constituted of one single image or a sequence of images may also influence on the choice of the applied method. Roughly speaking, methods that perform well for images could be less robust when dealing with videos, the same is also true for single viewpoint and multi-view data.

In this chapter, we explore the main concepts related to HAR by introducing some important definitions in section 2.2. Following this, we provide the main HAR application fields in section 2.3. Then, section 2.4 is devoted to present various types of human activities while section 2.5 summarizes the body parts involved in carrying out such movements. We

discuss in section 2.6 and section 2.7, the nature of the recorded data and the viewpoint of the acquisition device respectively. Similar to (19), we categorize HAR into contact-based and vision-based techniques in section 2.8 and we discuss drawbacks ans advantages of both categories. We introduce a brief description of the different validation means, where we categorize benchmark datasets according to the activity type and enumerate some of the metrics used for evaluating performances of the HAR systems in section 2.9. Finally, we summarize the chapter in Section 2.10.

2.2 Definitions

In this section, we introduce some definitions of main concepts related to HAR. These definitions may help to have a global idea about the studied field.

2.2.1 Activity

A human activity (HA) refers to the movement (s) of one or several parts of the person's body. This can be either atomic or composed of many primitive actions performed in some sequential order. It is to note that we use here the terminology of both action and activity to refer to the same concept.

2.2.2 Human Activity Recognition

Human activity recognition (HAR) refers to the process of identification and categorization of a sequence of recorded data from ubiquitous or visual sensors into well-defined basic activities (1). More specifically, human activity recognition should allow labeling the same activity with the same label even when performed by different persons under different conditions or styles. HAR systems attempt to automatically analyze and recognize such HAs using the acquired information from the various types of sensors (2, 3). Besides, HAR outputs can be employed to guide subsequent decision-support systems. For instance, authors in (38) proposed an HAR system that can help a teacher to control a multi-screen and multi-touch teaching tool, such as sweeping right or left to access the previous or next slide, call the eraser tool to rub out the wrong content, among others. Similarly, the work of (8) aims at ensuring a good implementation of various Human Computer interaction systems.

2.2.3 Activity Detection vs Activity classification

Human Activity Recognition systems are generally preceded by an activity detection task. This consists of the temporal identification and localization of such activity in the scene in a way to boost the understanding of the ongoing event. Therefore, the activity recognition task

can be divided into: classification and detection. The categorization of an activity, known as activity classification, consists in distinguishing the nature of the person movements using some spatial and temporal cues or any other meaningful features that best describe the ongoing action and assign it to its corresponding class.

2.2.4 Abnormal Human Activity

According to (39), an abnormal activity is defined as any out-of-ordinary and non-usual activity that may expose a person or group of people to danger in a particular context. An activity is considered abnormal when it is atypical and consists of undesirable acts. It is also an activity that occurs rarely and has not been expected in advance (40).

2.3 HAR Applications

HAR has been widely used for various applications such as human-computer interaction systems HCI, computer vision and augmented reality (5, 8, 9, 11, 41, 42, 43). It is important to stress the interest of having interactive and natural interfaces, where the user can use his performed actions to provide instructions to the machine. For instance, a speaker can control the presentation of the slides with the movements of his hand (19). Likewise, the recognition of human activity from static images or video sequences has several potential applications in many fields. Examples of human activity recognition applications include monitoring and evaluation of processes in industry as well as machines and devices control (8, 44, 45), fraud detection (2, 46), extraction of information from videos (3, 47), video assistance and surveillance (2, 3, 43, 45, 46, 47) and public security (7) where crowds' movements are tracked to detect violent or criminal situations. Given the increasing involvement of robots in our life, it is essential to equip them with the ability to understand intentions, emotions and behaviors of individuals. Thus, Robotics and video games have also taken benefits from the progress made in HAR systems (2, 3, 43, 46, 48). Finally, other applications of HAR systems touch in medical environments to ensure surgical operations or patient monitoring, interpretation of language signs (2, 46, 49, 50, 51, 52) as well as supervision of medication (43). For instance, a combination feature extraction method based on human activities recognition is introduced in (52) making possible to classify static signs of the sign language.

The interest in the development of these human activity-based applications can be justified by the fact that they provide very valuable and useful means of communication. However, the progress of the research in this field is also affected by the considerable changes in the technology trend and overall ecosystems.

2.4 Activities type

Human activities are regarded to be the means of communication between individuals, interactions with machines and with the environment in which we live. As mentioned earlier in this chapter, activities refer to body parts or whole body movements and is composed of several elementary actions performed in a temporal sequential order. They can be accomplished by one person or a group of people. We present in this section, a hierarchy of human activities depending on their complexity scaling from simple action to more complex events. Figure 2.1 shows this hierarchy.

1. *Elementary human actions*: This consists of simple atomic activities which indicate voluntary and/or intentional body movements that form the basis for building other more complex actions such as "raising the left hand" or "walking". They are easy to recognize and have been a center of interest of several research tasks like (45, 46, 47, 53).

2. *Gestures*: Typically, a gesture is a language or part of the non-verbal communication which can be employed to express significant ideas or orders. The gestures are a second type of activities which may be conscious like "applauding", and unconscious like "hiding the face with hands when getting shy" or "pulling out the hand when touching a hot material". Some gestures are universal, whereas others are related to quite specific social and cultural contexts. Among the works that are interested in gestures recognition, we can mention (8, 46, 54, 55).

3. *Behaviors*: These describe the set of physical actions and reactions of individuals in specific situations that are observable from the outside and which are relative to their emotions and psychological states. Proposals in (7, 43, 46) are examples of approaches that attempt to recognize human behaviors.

4. *Interactions*: These are reciprocal actions or exchanges between two entities or more, which modify the behavior of individuals or objects involved in the interaction. They are, in general, complex activities of two types: human to human such as "kissing" or human to object such as "cooking" which involves various kitchen utensils. The works presented in (55, 56) focus on the recognition of interactions.

5. *Group actions*: constitute the activities carried out by a group of people like "cuddling". These activities are more or less complex and difficult to track or recognize. The approaches suggested in (46, 55) make it possible to recognize complex activities.

Figure 2.1: Human activity types scaling from simple action to event

6. *Events*: These are human activities taking place in a specific environment or high-level activities which represent social actions between individuals such as "weddings and parties" (57, 58).

7. *Activities of daily living*: ADL can be defined as activities that people are used to do every day without any assistance. In general, they are complex activities, composed of many simple activities performed in an indoor environment. This particular type of activities related to daily living (ADL), has gained a particular attention in the computer vision community due to its importance in surveillance, assistance and patient monitoring environments. Understanding such activities is important because, in addition to providing information on the person's autonomy and ability to independent living, it provides personal safety to older people. Among research interested to activities of daily living, we can quote (43, 44, 59, 60, 61).

2.5 Body parts used for HAR

Human activities can be performed using only one part or several parts of the body, and are interpreted differently according to the culture of the region. Recognition of human activities may require analyzing the movements of different body parts of the individual, e.g., hands, feet, head, …etc (see Figure 2.2). The movements of a single limb or several at once enable to describe and give significant information on the underlined action. The hand, for example, can be tracked to detect the communication between individuals. The works presented in (8, 10, 46, 55) are interested in the recognition of hand gestures in particular. The foot is also a part which can be tracked to detect shifting and movements of people or other actions like walking, running, …etc. The majority of studies focus on tracking the full body performed activities. Authors of (3, 45, 46, 47, 53, 62) conduct their studies on the recognition of postures and human actions through tracking the whole body. Other works, such as (63, 64), focus on the follow-up of facial expressions to interpret specific types of human activities,

Figure 2.2: Different human body parts used to perform actions

especially in the case of handicapped or disabled people who can move neither their hands nor other parts of their bodies.

2.6 Image input versus video input

The research in human activity recognition can be classified, according to the nature of input data, into two sub-classes:

1. *Human activities recognition based on static images*: where the system can recognize activities from images like: sitting, walking, eating, ...etc. The activity is distinguishable compared to others by its characteristics. Authors of (8, 53, 65, 66) are interested to recognize human activities from still images;

2. *Human activities recognition based on videos*: Some activities cannot be recognized using solely a single current static image. There is a need to have access to extra information related to prior and post event occurring, through, for instance, examining previous and next frames. In this case, videos are more accurate, where the relation between two successive frames can be established. Works in (2, 45, 46, 47, 55, 62) attempt to recognize activities from videos using such approach.

2.7 Single viewpoint versus multi-view acquisition

The point of view is the place where the grabbing camera is located. Many researchers are interested to recognize activities from only one device placed in a suitable position while

others consider many devices placed at different positions to have different views of the activity performed. We can classify works of the state of the art into two categories:

1. *Single view acquisition*: This considers only one viewpoint because of the nature of the acquisition device which captures only the front part as in (8, 53, 55, 62). We give an example of single view recorded frames from the MSR Daily Activity 3D dataset (67) in Figure 2.3 a.

2. *Multiple view acquisition*: In the sequel, multiple devices are used to cover multiple views of the scene as in (24, 45, 68, 69). This allows to get extra information on the activity performed. We give an example of multi-view recorded frames from the Caviar dataset in Figure 2.3 b.

(a) **(b)**

Figure 2.3: Viewpoint of the acquisition device: (a) Samples from a single view dataset "MSR Daily Activity 3D (67)", (b) Samples from a multi-view dataset "the Caviar dataset"

2.8 HAR from contact-based to remote methods

HAR systems implementation is guided by two main streams of human computer interaction technologies: (1) Contact-based and, (2) Remote methods (54). Contact-based systems require the physical interaction of the user with the command acquisition machine or device (70). These methods are also straightforwardly impacted by the nature of data issued from the various sensory modalities and sources, e.g., accelerometers, multi-touch screens, body-mounted sensors or wearable sensors such as data gloves to analyze the human behavior. Nevertheless, contact-based systems are more and more abandoned because the physical

contact requires some skills and sophisticated equipment that make them accessible only to experimented users. Besides, in order to enable implementation in real world application scenarios, wearable sensors need as well to be easy, efficient, of adequate size, and should benefit from user's acceptability and willingness to perform continuous monitoring tasks. Among the currently developed HAR, one shall distinguish the vision-based (Remote methods). The latter attempts to simplify the human computer interaction task by allowing the human to use natural and intuitive manner in communication (54). In fact, human activities enable a user to convey ideas or thoughts through his gestures or combination of such activities (8, 54, 55). Intuitively, since vision-based systems use captured images or recorded video sequences to recognize activities, this can provide an edge to alternative approaches to win societal trust. Unlike contact-based activity recognition systems, vision-based systems do not require ordinary users to wear several and uncomfortable devices on different parts of their body. So, the "non-intrusive" character of these last systems allowed them to gain acceptability of use among the society.

2.9 Validation means

The validation of an approach is a vital step because it allows confirming by tangible proofs that its results are conforming and satisfying the requirements specified in the relative solution. In this context, we classify the research works according to the means used to check the proposals. To validate their approach, the authors of (45) used an experimental platform called DOMUS (71) which is designed and implemented by the MULTICOM team of the Informatics Laboratory of Grenoble. This functional apartment of 34 m^2 is equipped with various types of sensors and actuators in order to act on the environment (lighting, shutters, security systems, heating, ventilation, audio-video control…). Real data from 21 people, who performed predefined scenarios of different actions, were collected. Moreover, the authors of (47) validate their proposal through a set of real experiments allowing them to test their method on three sets of actions: MSR 3D Action (72), UT Kinect Action (73) and Florence 3D Action (73). For each set, half of the subjects was used for the training and the other half for the test. The authors of (42) have used the same validation means in order to support their proposal. However, they have used UT-interaction dataset (57) and UCF 50 dataset (58). The authors of (74) have conducted their experiments on five benchmark datasets: MSR Action 3D (72), UT Kinect (73), MSR C12 (75), Multiview 3D Action dataset (76) and NTU RGB+D dataset (77). The authors of (62) as well led experiments on the data extracted from the KTH database to validate and analyze the performance of their approach. They compared the results obtained by the naive Bayesian and SVM classifiers. In addition,

(8) created a dataset of actions performed by the members of the laboratory using a Web-Cam in order to evaluate the performance of their suggested model. Authors in (78) used the Kinect sensor in order to build a learning model of 22 human postures. These postures are divided into three categories: postures of the hand, foot and full body. Similarly, (8) executed the designed system on a dataset captured by Kinect, containing 1000 depth maps of hand gestures of 10 subjects with 10 catches for each gesture.

In addition, most of the works such as (42, 62) compare the results of their approaches with the results of other methods and approaches suggested in the literature to prove their effectiveness.

2.9.1 Open datasets

There exist many public datasets that can be used by researchers in order to validate their proposals and to evaluate their performance. According to (24), these datasets / databases can be grouped into several classes depending on the types of action they contain, the viewpoint as well as the nature of data: databases relating to movie scenes, social networks, human behaviors, human poses, atomic actions or daily life activities. Authors of (3) enumerate 13 sets of data captured using Kinect that can be used for training and testing. We quote the most used datasets in the literature and categorize them according to activity types. We consider in this classification only four types (levels): atomic action level, behavior level, interaction level and group activities level.

1. *Action level datasets*

   (a) *KTH Human Action Dataset*: It was created in 2004 by the Royal Institute of Technology of Sweden (79). It compromises 2391 sequences of six human action classes (walking, jogging, running, boxing, hand waving and hand clapping) performed several times by 25 subjects in four different scenarios. All sequences have a length of 4 seconds in average and were taken over homogeneous backgrounds with a static camera (Figure 2.4a).

   (b) *Weizmann Human Action Dataset*: It was created by the Weizmann Institute of Science in 2005 (80). It compromises 90 video sequences of 9 different people performing 10 simple actions (running, walking, skipping, jumping-jack, jumping forward on two legs, jumping in place on two legs, gallop-sideways, waving with two hands, waving with one hand and bending) (Figure 2.4b).

   (c) *Stanford 40 Actions dataset*: It was created by the Stanford vision Lab (81). It contains 9532 images of 40 different classes of actions.

(d) *IXMAS dataset*: It is a multi-view for view-invariant human action recognition dataset which was created in 2006 (82). It contains videos of 11 actors performing 13 daily life motions, 3 times each. These actions are recorded with 5 calibrated cameras from different views and include: crossing arms, stretching head, sitting down, …etc. (Figure 2.4c).

(e) *MSR Action 3D*: It was created by Wanqing Li in the Microsoft Research Redmond (72). It contains 567 depth map sequences of 10 subjects performing 20 action types twice or 3 times. The sequences were recorded using a Kinect device (Figure 2.4d).

2. *Behavior level datasets*:

(a) *VISOR dataset*: It was created in 2005 by the Imagelab Laboratory of the University of Modena and Reggio Emilia (83). It is composed of several types of videos sorted in different categories. The category "Videos for human action recognition in video surveillance" is used for human action and activity recognition and it contains 130 video sequences (Figure 2.5a).

(b) *Caviar dataset*: It was created in 2004. It is composed of two sets: the first one is filmed with a wide-angle camera lens in the entrance lobby of the INRIA Labs in Grenoble, France and the second set also uses a wide-angle lens along and across the hallway in a shopping center in Lisbon. This dataset compromises a number of videos of people performing 9 activities in two different places (Figure 2.5b).

(c) *Multi-Camera Action Dataset (MCAD)*: was created in the National University of Singapore (84). It was designed to evaluate the open-view classification problem under surveillance environment. 18 daily actions which are inherited from the KTH, IXMAS and TRECIVD datasets are recorded using 5 cameras and performed by 20 subjects. For each camera, each action is produced by a subject 8 times (4 times during the day and 4 times in the evening).

3. *Interaction level datasets*:

(a) *MSR Daily Activity 3D Dataset*: It was created by Jiang Wang in the Microsoft Research Redmond (67). It consists of 320 sequences for each channel: depth maps, skeleton joint positions and RGB video of 10 subjects performing 16 activities such as drinking, eating, reading …etc. Each activity is carried out twice, once in standing position and once in sitting position (Figure 2.6a).

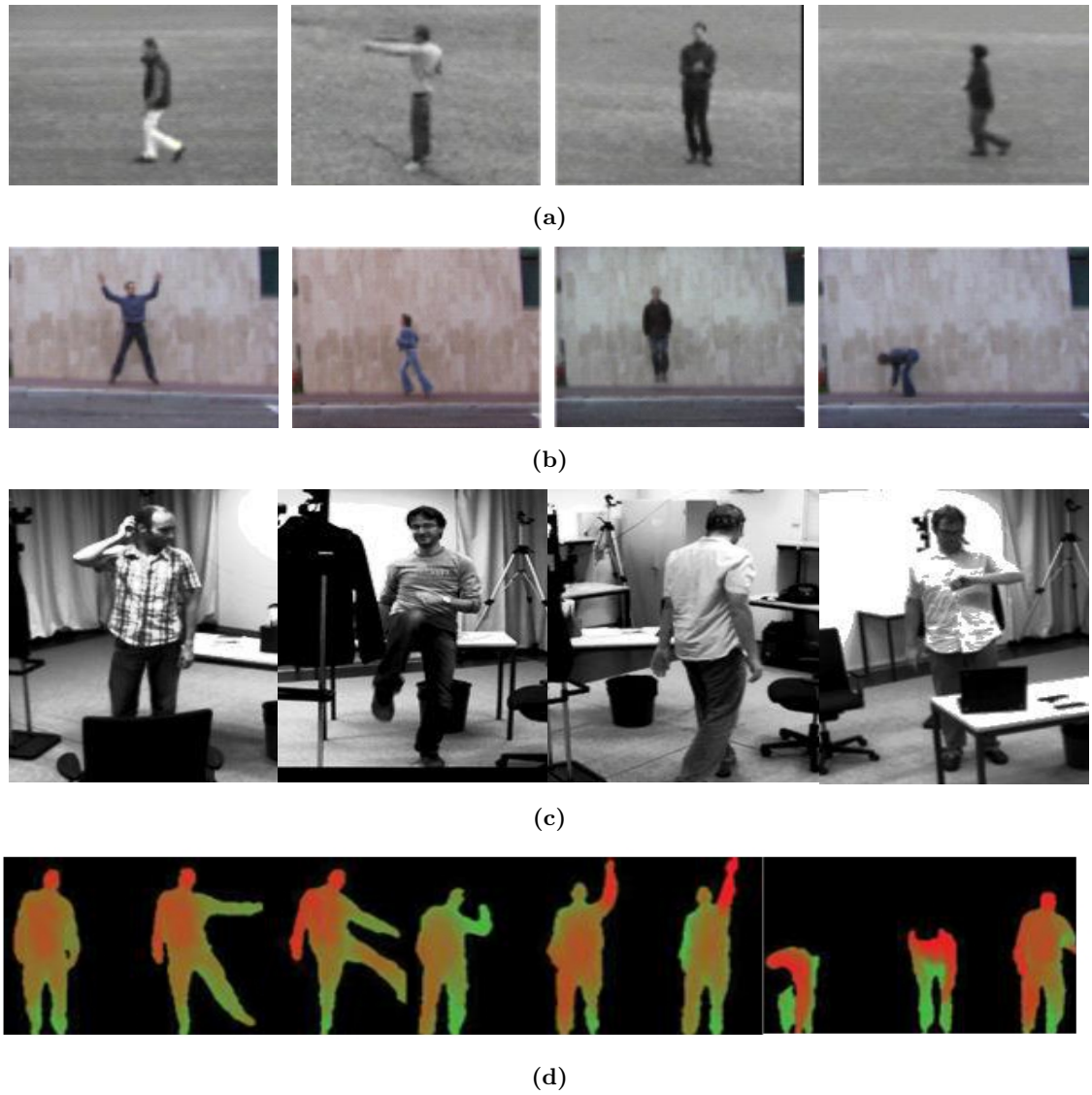


Figure 2.4: Samples from action level datasets: (a) KTH Human Action Dataset (79), (b) Weizmann Human Action Dataset (80), (c) IXMAS dataset (82), (d) MSR Action 3D dataset (72)

(a)

(b)

Figure 2.5: Samples from behavior level datasets: (a) Visor Dataset (83), (b) Caviar dataset

(b) *50 Salads dataset*: It was created at the University of Dundee (85). It is composed of video sequences of 25 people preparing 2 mixed salads, having a total duration of 4 hours.

(c) *MuHAVI dataset or Multicamera Human Action Video Dataset*: It was created in 2010 by the Faculty of Science, Engineering and Computing of Kingston University. It aims at evaluating silhouette-based human action recognition methods. It consists of videos of 17 action classes performed by 14 actors several times. The data were recorded using 8 non-synchronized cameras located on 4 sides and 4 corners of a rectangular platform (Figure 2.6b).

(d) *UCF50*: It was created by the center for research in Computer Vision, University of Central Florida, USA in 2012 (58). It is composed of 50 action categories, collected from realistic YouTube videos. This dataset is an extension of YouTube Action dataset (UCF11) which has 11 action categories.

(e) *UCF Sports Action Dataset*: It was created by the Center for Research in Computer Vision, University of central florida, USA in 2008 (86, 87). It is composed of 150 sequences of 11 action categories collected from various sports broadcasted on television channels.

(f) *ETISEO dataset*: It was created in 2005 by the INRIA Institute (88). It aims at improving video surveillance algorithms. It provides videos of people carrying out several activities in 5 different scenarios: apron, building corridor, building entrance, metro and road.

(g) *Olympic Sports Dataset*: It was created in 2010 by the Stanford Vision Lab (89). It contains 50 videos of athletes practicing 16 different sports. All video sequences are gathered from YouTube.

(h) *UT-Interaction dataset*: It was created by the University of Texas (90) within the Contest on Semantic Description of Human Activities (SDHA), a research competition to recognize human activities in realistic scenarios, which was held in conjunction with the 20th International Conference on Pattern Recognition (ICPR 2010). It contains 20 video sequences of continuous executions of 6 classes of human-human interactions. Several participants with more than 15 different clothing conditions appear in the videos (Figure 2.6c).

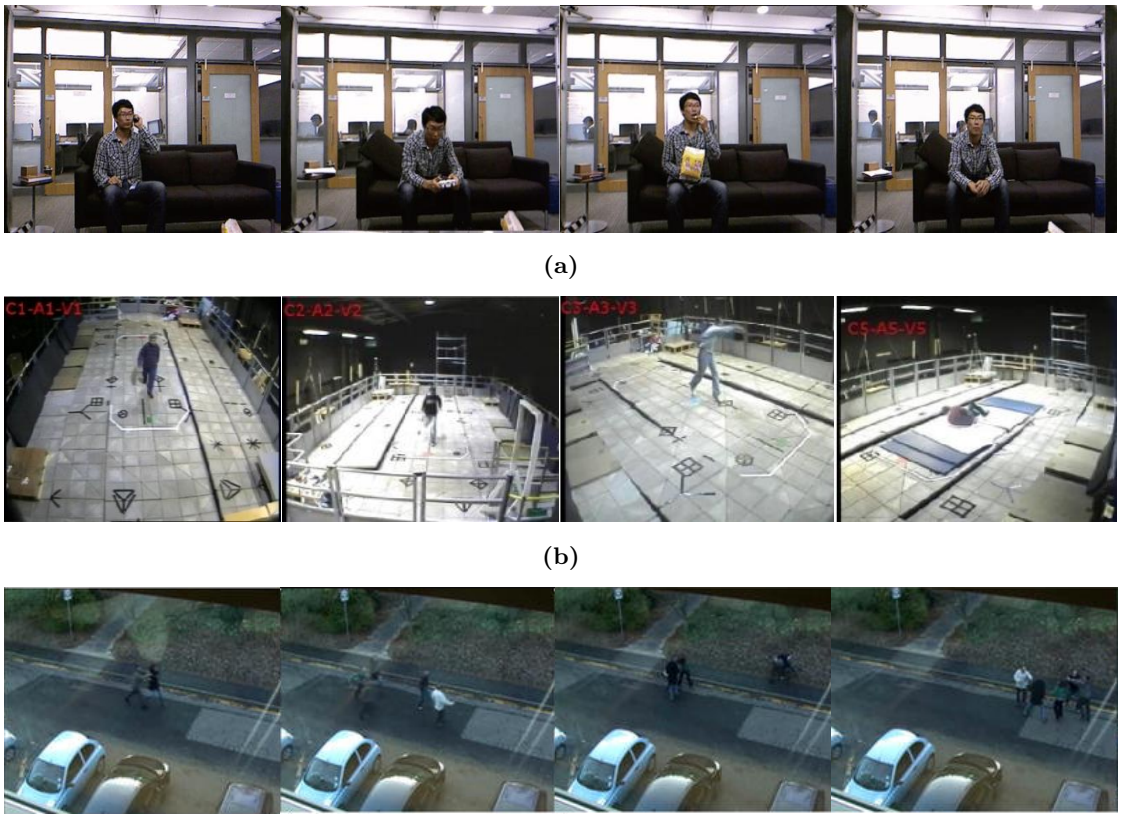
(i) *UT-Tower dataset*: It was created as well in the same context of UT-interaction dataset (91). It consists of 108 video sequences of 9 types of actions. Each action was performed 12 times by 6 individuals. The dataset is composed of two types of scenes: concrete square and lawn.

4. *Group activities level datasets*:

(a) *ActivityNet Dataset*: It was created in 2015 (92). It consists of 849 video hours that illustrate 203 activity classes with 137 untrimmed videos for each activity class. It encompasses three scenarios to compare human activity understanding algorithms: untrimmed video classification, trimmed activity classification and activity detection. It is considered as a large-scale video dataset covering a wide range of complex human activities (Figure 2.7a).

(b) *The Kinetics Human Action Video Dataset*: It was created by the DeepMind team in 2017 (93). The initial release (Kinetics_400) contains 400 human action classes with at least 400 video clips for each action taken from different YouTube videos. Kinetics_600 is an approximate super-set of the initial Kinetics_400 dataset that covers 600 human action classes with at least 600 video clips for each action class. It consists of approximately 500,000 video clips, and each clip lasts around 10 seconds and is labeled with a single class. It is a large-scale, high-quality dataset of YouTube video URLs which include a diverse range of human focused actions.

(c) *HMDB-51 dataset*: It was created in 2011 by the Serre Lab, Brown university USA (94). It consists of 6849 clips of 51 action categories collected from various sources (movies, public data-bases such as Prelinger archive, YouTube and Google videos). It is considered as one of the largest datasets of human activities recognition.



(a)

(b)

(c)

Figure 2.6: Samples from interaction level datasets: (a) MSR Daily Activity 3D dataset (67), (b) MuHAVI dataset, (c) UT-Interaction dataset (90)

(d) *Hollywood dataset*: It was created in 2008 by INRIA (Institut national de recherche en informatique et en automatique), France (95). It is composed of short sequences of human actions collected from realistic videos retrieved from 32 movies.

(e) *Hollywood2 dataset*: It was created as well by INRIA in 2009 (96). It was proposed to provide realistic and challenging settings (multiple persons, cluttered background …). It is composed of 3669 video clips of 12 classes of human actions and 10 classes of scenes collected from 69 movies.

(f) *UCF-101 Action Recognition Dataset*: It was created by the Centre for Research in Computer Vision, University of Central Florida, USA in 2012 (97). It is an extension of UCF50 dataset (58) which has 50 action categories. It is composed of 13 320 videos of 101 realistic action categories collected from YouTube. It gives the largest diversity in terms of actions and realistic settings (viewpoint, illumination conditions …etc.) (Figure 2.7b).

(g) *YouTube Action Dataset*: It was developed by the Center for Research in Computer Vision, University of Central Florida, USA in 2009 (98). It contains 11 action categories and it is very challenging due to large variations in camera motion, viewpoints, illumination condition, …etc.

(h) *Behave dataset*: It was created in 2004 by the School of Informatics of Edinburgh University. It aims at detecting unusual human activities. It is composed of two sets: optical flow data and multi-agent interaction data. The first set is composed of 30 optical flow sequences from the Waverly train station while the second one comprises two views of various scenarios of people interactions (Figure 2.7c).

(i) *Video Web Dataset*: It was created in 2010 by the Video Computing Group, belonging to the Department of Electrical Engineering at the University of California Riverside (UCR) (99). It consists of 2.5 hours of videos of 10 actors interacting with each other, with vehicles or with facilities. Each video is recorded using a camera network of minimum of 4 and maximum of 8 cameras.

(a)

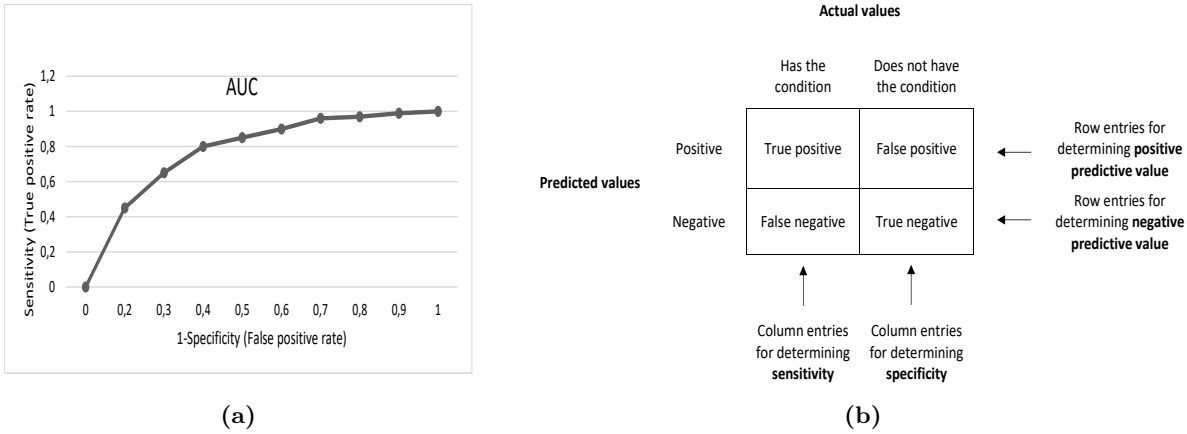
(b)

(c)

Figure 2.7: Samples from group activities level datasets: (a) ActivityNet Dataset (92), (b) UCF-101 Action Recognition Dataset (97), (c) Behave dataset

Table 2.1: Classification of Benchmark datasets based on activity types

Dataset	Action	Behavior	Human-object interaction	Human-human interaction	Group activities
KTH	X				
Weizmann	X				
Stanford 40	X				
IXMAS	X				
MSR Action 3D	X				
VISOR		X			
Caviar		X			
MCAD		X			
MSR Daily Activity 3D	X		X		
50 Salads	X		X		
MuHAVI	X	X	X		
UCF50	X		X		
UCF Sports	X		X		
ETISEO			X	X	
Olympic Sports	X		X		
UT-Interaction			X	X	
UT-Tower			X	X	
ActivityNet	X	X	X	X	X
Kinetics	X	X	X	X	X
HMDB-51	X	X	X	X	X
Hollywood	X	X	X	X	X
Hollywood2	X	X	X	X	X
UCF-101	X	X	X	X	X
YouTube Action	X	X	X	X	X
Behave				X	X
Video Web			X	X	X



(a) **(b)**

Figure 2.8: Examples of performance evaluation metrics (AUC and confusion matrix): (a) Diagram demonstrating how to calculate the AUC, (b) Structure of the confusion matrix

For instance, the authors of (62) use the KTH dataset (79), which contains grey level video sequences of low-resolution (160 x 120 pixels), representing atomic human actions. The sets of test and training are separated and contain 461 and 328 images respectively. The authors of (47) use three datasets to test the performance of their approach; 3D MSR Action (72), UT Kinect (73) and 3D Florence Actions (100). The two last datasets are captured using a fixed Kinect sensor, which is composed of 10 and 9 actions performed by 10 different subjects with a total of 199 and 215 sequences of actions respectively. The authors of (45) recorded video sequences of several activities (dressing, shopping, cleaning, listening to the radio …) performed by 21 persons in predefined scenarios. Moreover, (78) used the Kinect sensor to build a training model for 22 human postures. These postures are divided into three categories: postures of the hand, foot, and full body. For each posture, 100 samples are used. In the same way, (8) have built a training dataset captured by Kinect, which contains 1000 depth maps of hand gestures of 10 subjects with 10 catches for each gesture. We present in Table 2.1, a classification of benchmark datasets based on activity types, in order to enable the reader to better select the suitable dataset for his study.

2.9.2 Evaluation metrics

Several performance metrics used in different classification fields have been adapted and used for human activities recognition. Based on (101), we quote in this section frequently used metrics such as accuracy, precision, recall, …etc. Before summarizing these metrics, we define firstly what would be True Positive, True Negative, False Positive and False Negative for action recognition (see Figure 2.8 b) as follows:

□ True Positive (TP) = actions where the actual and predicted transactions are correct.

- False Negative (FN) = actions which belong to a particular class and are actually predicted to be not from this class.
- True Negative (TN) = actions where the actual and predicted transactions do not correspond to the searched class.
- False Positive (FP) = actions where the actual transactions do not correspond to the searched class, but predicted to be from the searched class.

We describe frequently used performance metrics in the following:

1. *Sensitivity*: It is also called true positive rate, recall or probability of detection. It corresponds to actual positive cases predicted as positive. For action recognition, sensitivity measures the proportion of activities predicted in their classes. Likewise, (1 - sensitivity) determines the failure of the system to detect actions. Mathematically, this can be expressed as:

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.1}$$

2. *Precision*: It is also called Positive Prediction Value (PPV) and corresponds to the likelihood of a detected instance of activity to its real occurrence. Likewise, (1 - precision) determines the probability of the recognizer incorrectly identifying a detected activity. Mathematically, this can be expressed as:

$$Precision = \frac{TP}{TP + FP} \tag{2.2}$$

3. *Specificity*: It is also called true negative rate or false positive rate (FPR). It corresponds to actual negative cases predicted as negative. It measures the system sensitivity to negative class. Mathematically, specificity can be calculated as follows:

$$Specificity = \frac{TN}{TN + FP} \tag{2.3}$$

4. *Negative Predictive Value (NPV)*: It is often referred to "negative precision" and measures the likelihood that a negative identification is correct relative to all negative identifications. Mathematically, this can be expressed as:

$$NPV = \frac{TN}{TN + FN} \tag{2.4}$$

5. *F_Measure*: It determines the harmonic mean of precision and recall. It gives information about the test's accuracy. Hence, F_measure determines at the same time, how precise and robust is the classifier. It reaches its best value at 1 and worst at 0. Mathematically, this can be expressed as:

$$F\_Measure = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2.5}$$

6. *Accuracy*: It measures the percentage of correct predictions relative to the total number of samples. The accuracy gives good results when the classes are equally sampled. Mathematically, this can be expressed as:

$$Accuracy = \frac{\text{Correct Predictions}}{\text{Total Predictions Made}} = \frac{TP + FN}{TP + FN + TN + FP} \tag{2.6}$$

7. *Likelihood Ratio*: computes the likelihood of an activity predicted when it matches the ground truth compared to the likelihood when it is predicted wrongly. It can be computed for both true positive and true negative results. Mathematically, this can be expressed as:

$$LR+ = \frac{Sensitivity}{1 - Specificity} \qquad LR- = \frac{1 - Sensitivity}{Specificity} \tag{2.7}$$

8. *Area Under Curve (AUC)*: It is widely used for binary classification problems. It measures the probability of the classifier to rank a positive randomly chosen sample higher than a negative randomly chosen sample. In perfect cases, its value is 1. AUC is the area under the curve of plot False Positive Rate (Specificity) vs True Positive Rate (Sensitivity) at different thresholds ranging in [0, 1] (see Figure 2.8 a).

9. *Confusion Matrix*: It is also called error matrix and gives a summary of prediction results and describes the complete performance of the model. The confusion matrix shows the errors being made by the classifier as well as their types. Each row of the matrix represents instances of predicted class and each column represents instances in an actual class or vice versa (see Figure 2.8 b).

10. *Intersection over Union (IoU)*: called also Jaccard index or Jaccard similarity coefficient. It measures the accuracy of the detector on a particular dataset. If we refer to area of overlap between predicted bounding box and ground-truth bounding box with "Area of overlap" and the area encompassed by both the predicted bounding box and the

ground-truth bounding box with "Area of Union", this ratio can be calculated as the following:

$$IoU = \frac{\text{Area of overlap}}{\text{Area of union}} \tag{2.8}$$

2.10 Conclusion

The need to understand and interpret effectively human activity has become unavoidable in several applications of computer vision, HCI, robotics, security and home monitoring. This chapter allowed us to give an overview of the HAR field. We initially introduced the HAR process, where we define some main concepts. Then, we discussed the different applications of HAR, and the major objectives intended by these systems. We also provided a taxonomy of human activity types and the involved body parts. Moreover, classification of HAR approaches according to the nature of data and the viewpoint of acquisition was also given. Next, we discussed the existing implementations of HAR systems: contact-based and remote methods and finally the means used in performance validation of such systems by providing the mostly used metrics and benchmark datasets. Before presenting our proposed framework for vision-based HAR, we scrutinize in the next chapter the existing methods in the literature by discussing their limitations and the challenges they are still facing till today.

3

Literature review: Vision-based Human Activity Recognition

3.1 Introduction

Human activity recognition has been studied significantly in the literature. In some previous works (3, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29), HAR systems have been extensively reviewed and discussed. However, summarizing all the existing HAR systems is quite hard. We will restrict in this section to review only vision-based HARs. We advocate different techniques introduced for HAR from various angles, where we include supervised, unsupervised, handcrafted features based, features learning based, uni-modal and multi-modal techniques. In line with our following overview that completes and updates the aforementioned existing surveys, it is important to highlight the limitations of the proposed techniques in the literature. This allows us to resolve in the next chapters some of the underlined challenges.

In this chapter, we review a plenty of vision-based human activity recognition studies and approaches. We first, discuss some of the related surveys that introduce the recent advances in human activity recognition topic in section 3.2. Then, we classify in section 3.3, vision-based HAR approaches according to various criteria: 1) feature extraction process that could rely on handcrafted features or learning methods; 2) recognition stages which are summarized into detection, tracking and classification; 3) source of input data which may be recorded using only one modality or is combined from different data modalities; 4) machine learning supervision level that includes supervised level, unsupervised level and semi-supervised level. Section 3.4 is devoted to limitations that face the development of reliable vision-based HAR systems while section 3.5 discusses the current challenges. Finally, we summarize the chapter in section 3.6.

3.2 Vision-based HAR Related surveys

The state-of-the-art methods of HAR task are studied and surveyed in different papers (1, 3, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 102, 103, 104, 105, 106, 107, 108, 109). These survey papers introduce the recent advances in automated human activity recognition topic. A few number of these works such as (21, 26, 27, 28, 105, 109, 110) provide a comprehensive review on different aspects of HAR methods, while most of them look at HAR task from a specific point of view. For example, (1) classifies HAR approaches according to the spatial-temporal characteristics of actions, video segmentation and recognition systems and camera modalities. Likewise, (106) advocates a taxonomy-based approach and compares the advantages and limitations of each method. The authors examined both simple human actions and high-level activities. In addition, the authors of (103) review HAR approaches according to the complexity of features involved in the action recognition process. Similarly, (104) categorizes human activities according to their complexity into action and activity and then reviews the major approaches for recognizing human actions and activities. Authors of (20) summarize existing methods for human activity recognition from still images and categorize them into two big categories according to the level of abstraction and the type of features each method uses. Otherwise, (29, 108) compare techniques related to image segmentation, feature extraction and activity classification, and then discuss advantages and limitations of each of them. Furthermore, the authors of (24) categorize HAR approaches into two broad categories: uni-modal and multi-modal with regards to the source channel each of these approaches employs for human activities recognition. They also reviewed the existing publicly available human activity datasets and examined the requirements for building both ideal HAR dataset and system.

On the other hand, (3) focused on techniques that use 3D and depth data, while (111, 112, 113) surveyed 3D skeleton-based human representation and action recognition approaches. Interestingly, (25) represents the semantic-based human recognition methods using still images and videos. It identifies semantic space and semantic-based features such as pose, poselet, related objects, attributes, and scene context. It also briefly discusses the potential applications of semantic approaches. Authors of (23, 107) provide a classification of common Kinect-based motion recognition approaches and review each approach accordingly. They highlight Microsoft Kinect sensor applications in various domains as well as the publicly available Kinect datasets.

Overviews of current progress in human activities recognition are also presented in (2, 46, 114, 115). These surveys analyze popular techniques used for object segmentation and activity recognition. The authors discussed as well the merits and demerits of such methods and proposed possible future scopes in this area of research. Authors in (116) were interested

into knowledge-based human activity recognition methodologies which are not well covered in the literature. More specifically, they survey methods and techniques used in the literature to represent and integrate knowledge and reasoning into the recognition process. These methods are categorized in terms of statistical, syntactic and description-based approaches.

On the other hand, reviews of vision-based hand gesture recognition methods were presented in (10, 50, 117, 118, 119, 120, 121, 122). The purpose of these reviews was to introduce the field of gesture recognition as a mechanism for interacting with computers and provide researchers with a summary of advances in hand gesture recognition to help identify areas where further research is needed. Authors in (10) focus on gesture taxonomies, their representations, recognition techniques, software platforms and frameworks as well as hand gesture recognition applications. Other researchers are interested in behavior and event understanding rather than actions and gestures. We can quote as examples (123, 124, 125, 126, 127, 128, 129). Authors in (125, 126) discuss the advantages and the drawbacks of existing approaches for behavior understanding as well as related available datasets. They also highlight open research challenges and several important future directions. (128) examines complex event recognition techniques. They identify a number of limitations with respect to the employed languages, probabilistic models and their performance as compared to the purely deterministic cases. Based on those limitations, promising directions for future work are then highlighted. Authors in (129) attempt to summarize techniques in understanding activities of a person and/or a group as well as their social interactions. For instance, understanding crowd behavior is deemed essential to surveillance and security purposes. Motivated by this fact, many surveys are devoted to crowd analysis such as (130, 131, 132, 133), in order to provide an overview of the major techniques applicable for classifying abnormal behavior in a crowded scene scenario. On the other hand, reviews on vision-based Ambient Assisted Living, patient monitoring like (60, 134, 135), fall detection and abnormal human activity recognition such as (4, 136) were proposed in the literature. Other researchers were interested in motion analysis to recognize human activities and many reviews were presented in this field. We can mention for instance (137, 138, 139, 140, 141, 142). Authors in (141) attempt to summarize human motion analysis algorithms that use depth imagery. (137) discusses three sub-topics of human motion analysis: human body parts-based motion analysis, moving human tracking and image sequences-based human recognition. (138) views human body motion as a hierarchical process with four steps: initialization, tracking, pose estimation and recognition. A comparative analysis of methods based on handcrafted representations and solutions that involve learning architectures is carried out in (49) and (143). In both surveys, the authors discuss recent advancement in human action representations alongside the associated pros and cons. Reviews on deep-learning based methods of human activities recognition were

provided in (144, 145, 146, 147). They analyzed the advantages and limitations of current existing techniques and discussed the potential directions for future research. Authors in (146) were interested particularly to RGB-D-based human motion recognition using deep learning architecture and focused on three architectures of neural networks: CNN, RNN and other structured networks. On the other hand, the authors of (144) enumerate a list of datasets in different complexity levels and compare the performance of deep learning-based approaches to other existing works.

Surveys on dataset benchmarks for human action recognition from visual data constitute another field of research tackled in (148, 149, 150, 151, 152, 153). They aim to guide researchers in the selection of the most suitable dataset for benchmarking their algorithms. These surveys also present the best performance scores achieved by various HAR methods on these benchmark dataset. The performance analysis includes the number of activity classes, complexity of events, application domain and impact of the ground truth. In addition, (150) presents a summary of the results obtained on the recent ASLAN benchmark (154), which was designed to reflect on the variety of challenges that modern activity recognition systems are expected to overcome. Authors in (152) propose a novel dataset, called CONVERSE, that represents complex conversational interactions between two individuals via 3D pose. Similarly, authors in (153, 155, 156, 157, 158) present a set of comprehensive reviews of the most commonly used RGB-D video-based activity recognition datasets. Relevant information in each category is extracted in order to help researchers to easily choose appropriate data for their needs. Moreover, the reviews highlight the evaluation protocols, and the limitations of the publicly available datasets. A guidance on future creation of datasets and establishment of standard evaluation protocols for specific purposes is also provided.

Authors in (101) examine another aspect of human activity recognition by analyzing several factors that influence the evaluation of activity recognition approaches. Especially, they reviewed many of the commonly used metrics, outlined the sources of errors in such systems and presented different methods for detecting and labeling these errors.

In this chapter, we discuss the most significant advances reported recently in the literature covering both the general aspects of human activities recognition and the specific vision-based HAR systems. For a comparison analysis of the abovementioned surveys refer to Table A.1 in the appendix.

3.3 Vision-based HAR approaches

HAR methods are composed of three important components: (1) Video frame segmentation for action detection, (2) Action representation with respect to posture and motion of the human body, and (3) Learning process that recognizes these actions. We categorize in this

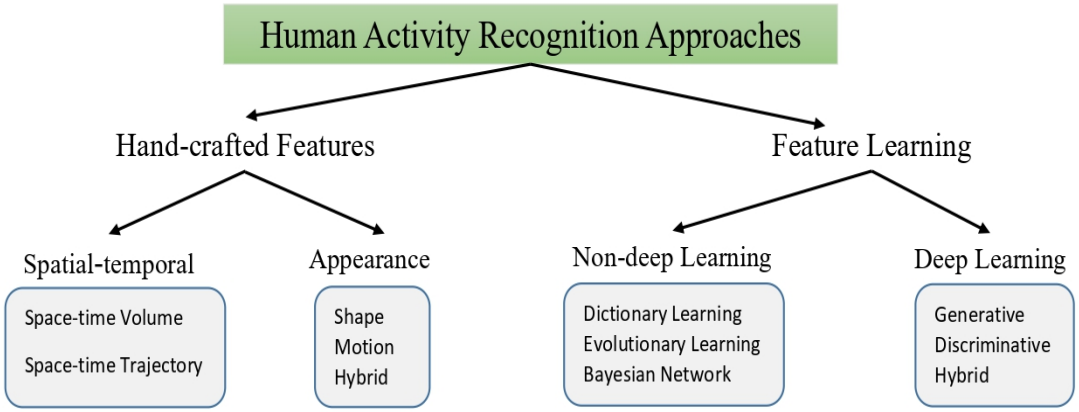


Figure 3.1: Vision-based Human activities recognition approaches

section, HAR approaches according to feature extraction process in the first subsection, to the three stages of recognition process in the second subsection, to the source modalities of input data in the third subsection and finally to machine learning supervision level in the fourth subsection.

3.3.1 HAR approaches according to feature extraction process

In this subsection, we present a classification of HAR approaches according to feature extraction process into handcrafted representation based and learning based approaches. Figure 3.1 summarizes the HAR methods as follows:

Methods based on Handcrafted features rely on human ingenuity and prior knowledge to extract discriminating features. These types of methods involve three major steps: (1) Foreground detection that corresponds to action segmentation, (2) Feature selection and extracting by an expert and (3) Classification of action represented by the extracted features (19). The input images or video frames are analyzed to extract the most significant features that are used, then to build the descriptor. The classification is performed using a generic trained classifier which makes this family of approaches low cost, flexible and does not rely on large sets of samples for training. Methods based on handcrafted features are: spatial-temporal-based approaches as discussed below, appearance-based approaches, and other methods like local binary pattern LBP which is a visual descriptor used for texture classification and fuzzy logic.

Human activities can be seen either static or dynamic. Static activities are described using the orientation and the position of the limb in space while dynamic activities are described as movements of these static activities (10). Hence, action recognition can be based either

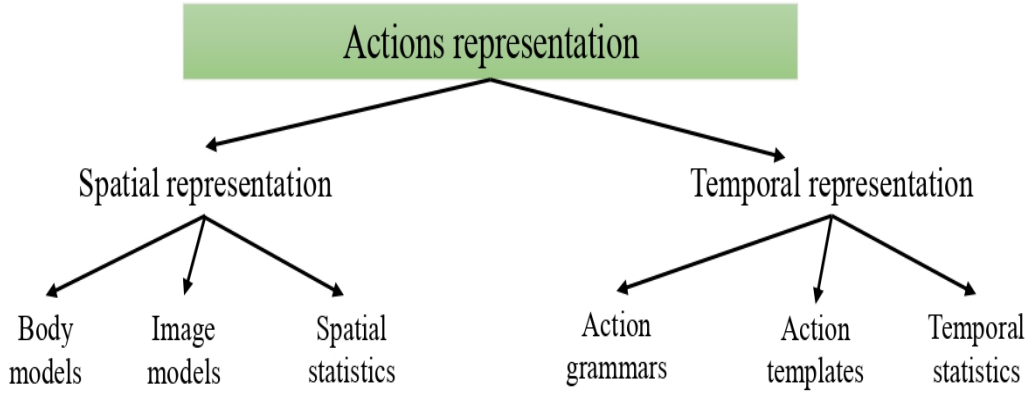


Figure 3.2: Spatial and temporal representations of actions

on spatial or temporal cues that describe and recognize these actions (1). Both spatial and temporal representations can be categorized into three classes (see Figure 3.2).

1. *Spatial representations*: This can also be categorized into three subclasses.

   (a) *Body models*: This allows recovering human body pose from features using the spatial structure of the action with reference to the human body. One potential method is the reconstruction of the 3D body model that aims at representing the body as a kinematic joint model (74, 159) and recognizing the underlined action using joint trajectories. An alternative method is the direct recognition from 2D models without employing 3D models.

   (b) *Image models*: This corresponds to a holistic representation of actions that use a regular grid bounded by a region of interest centred around the person to detect and compute features. For instance, we can quote silhouettes (160, 161, 162), contours (160), motion history images (MHI) (163, 164), motion energy images (MEI) (163, 164) and optical flow (165, 166, 167, 168) that are used to describe actions and movements.

   (c) *Spatial statistics*: This enables us to represent local actions through a set of statistics of local features from the surrounding regions. These regions are obtained after either dense or sparse decomposition of the underlined image or video. For instance, we can mention statistical methods and space-time interest points (163, 169, 170) which calculate Spatial-Temporal Interest Points (STIP) of the image and assign each region to a set of features, to provide spatial distribution

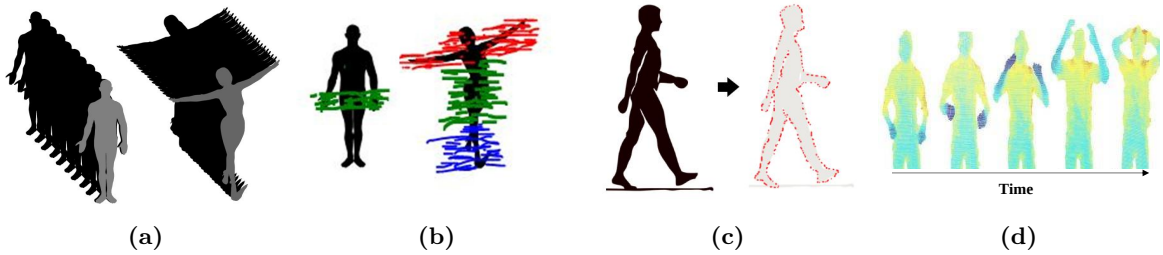


Figure 3.3: Examples of handcrafted feature extraction approaches: (a) Space-time volumes, (b) Space-time trajectories,(c) Shape-based methods: Contour features,(d) Motion-based methods (174).

of local image features. Methods based on spatial-temporal can be classified as follows:

- ☐ *Volume-based*: These approaches represent video as spatial-temporal volume and may rely on features like texture, color, posture, histograms of optical flow, histograms of oriented gradients …etc. Actions are recognized using similarity between the two volumes. Volume-based approaches can not work efficiently when the scene is crowded, they are only suitable for simple action or gesture recognition. For example, Scale Invariant Feature Transformation (SIFT) is used as a 2D interest point detector in (171) while corners and Laplacian of Gaussian are used in (172) to detect 3D interest points. Figure 3.3 a shows a representation of the human body with space-time volumes based methods.
- ☐ *Trajectory based approaches*: This family of approaches represent joint positions of the body with 2D or 3D points which are further tracked along the video to compute action trajectories. These tracked changes in the posture are used to construct the 3D representation of the activity which is considered to be a set of spatial-temporal trajectories. These methods are powerful against noise, view and/or illumination changes, and are useful for recognizing complex activities. For instance, (173) calculates several descriptors (HOG, HOF and motion boundary histogram MBH) and trajectories by tracking densely sampled points using optical flow. Figure 3.3 b shows a representation of the human body using spatial-temporal trajectories based methods.

2. *Temporal representations*:

   (a) *Action grammars*: they represent the action as a sequence of moments. Each moment is described by its own appearance and dynamics. To perform action

recognition task, features are grouped into similar configurations called states and temporal transition between these states are learned (175). Hidden Markov Models (176, 177), CRF (178, 179), regressive models (180, 181, 182) and context-free grammars (183, 184) are examples of action grammars.

(b) *Action templates*: This consists of a set of temporal representations that allow representing the appearance of temporal blocks of features and dynamics called templates (164, 185). Templates are computed over long sequences of frames and dynamics are represented using features of several frames loaded in feature vector or in space time volume. Typically, Fourier Transform (47, 186, 187), Wavelet representations (188, 189) and trajectories of body parts (190, 191) are templates that can be used for action recognition.

(c) *Temporal statistics*: They are based on statistical models and are as well temporal representation of action. Statistical models (192) are used to describe the distributions of unstructured features over time. These features represent the appearance of actions.

3. *Appearance based approaches*: They use 2D or 3D depth images and are based on shape features, motion features or any combination of both features. These methods have been significantly simplified due to the emergence of depth cameras. A quick advance is felt in the skeleton-based recognition approaches as well with the advent of depth sensors and algorithms of real-time skeleton estimation (51). According to (47), they can be classified into two categories:

   - ☐ Joint locations, which consider the skeleton as a set of points.
   - ☐ Joint angles, which assume the human body as a system of rigid connected segments and the movement as an evolution of their spatial configuration (2, 193, 194).

   As an appearance based approach, (47) proposes an efficient skeleton representation for activity recognition based on the body parts. The authors model the 3D geometric relationships between the body parts using rotations and translations. The work proposed in (78) is based on the Microsoft Kinect sensor and presents a new method of HAR while using the machine training. Actually, the growing interest in the recognition, generated by the release of Kinect, is due to the fact that the skeleton information can be deduced from depth images (53). A transparent standardized input interface, making it possible to reduce HCI constraints was proposed in (55). It is used to imitate in real time the motions carried out by the user, using an articulated 3D model. The system is based on the image analysis to detect the trajectory of the hand, to determine

its configurations and interpret them like 3D postures. This solution is very fast and computationally efficient, and can be used as a control entry of a mobile robot.

Generally, the appearance based approaches can be classified according to either shape or motion based characteristics.

(a) *Shape based methods*: They capture local shape features such as contour points, local region, silhouette and geometric features from the human image or video using foreground segmentation. For example, (8) proposes a process for hand gesture recognition in the 3D point cloud, which explicitly uses 3D information of depth maps where the color information is ignored. It submits a standardized descriptor of features, making it possible to effectively represent the various positions of the hand. Figure 3.3 c shows a representation of the human body using contour points.

(b) *Motion based methods*: They include optical flow and motion history volume, which are then used for action representation (82, 166, 167). Then, a generic classifier on top of this representation is implemented for action recognition. The research of (62) proposes to recognize actions using vector quantization of motion descriptors. This method is related to a meta-algorithm combining the histograms of optical flow and classifiers of bag-of-words. An example of action recognition proposed by (174) is given in Figure 3.3 d.

(c) *Hybrid methods*: they combine shape and motion features to represent actions (68, 160, 195, 196, 197).

Recently, feature learning (198, 199) has started to become popular in different computer vision applications such as pedestrian detection (200, 201), image classification (202, 203), vision-based anomaly detection (204), …etc. Besides, many of the learning-based action recognition approaches rely on the end-to-end learning which consists in transformations from pixel-level to action classes. The feature-learning based methods for the HAR task can further be grouped into (1) Traditional and, (2) Deep-learning-based methods.

1. *Traditional approaches*: This includes methods like genetic programming (205, 206), dictionary learning (207, 208) and Bayesian networks (209, 210).

   (a) *Dictionary learning*: This provides a sparse representation of the input data by a linear combination of basis dictionary atoms. They use an end-to-end unsupervised learning to learn the dictionary and the corresponding classifier within a single learning procedure. The concept of these methods is similar to visual

Bag of Words model (BoW) that generates global data representations. For instance, (211) presents a weakly-supervised cross-domain dictionary learning approach to adapt knowledge of one action dataset to another action dataset. The proposed method learns a reconstructive, discriminating and domain-adaptive dictionary pair and the corresponding classifier parameters without using any prior information. As aforementioned, methods based on spatial-temporal features are very popular, and have achieved state-of-the-art results on activities classification task. However, the use of over-complete dictionaries proves to be more interesting because they can produce even more compact representations. Therefore, the authors of (42) have combined these two concepts to propose a solution of HAR from multi-part missing video.

(b) *Genetic programming (GP)*: GP is an evolutionary method inspired by the process of biological evolution and its fundamental mechanism (19). The principle of genetic programming is to search a space of possible solutions without having any prior knowledge and it allows discovering functional relationships between features in data enabling its classification. Genetic programming has been used to construct holistic descriptors that allow to maximize the performance of action recognition tasks (205). In (205), the authors developed an adaptive learning methodology using GP to evolve discriminating spatial-temporal representations, which simultaneously fuse the color and motion information, for high-level action recognition tasks.

(c) *Bayesian networks*: These methods are probabilistic graphical models that infer the conditional dependencies using directed acyclic graphs (19). The graph nodes and edges represent the random variables and their associated conditional dependencies, respectively. The probability computations are performed using the Bayesian inference. In (209), the authors adopted a Bayesian Network (BN) to represent and capture the semantic relationships among action units, as well as the correlations of the action unit intensities, to more accurately and robustly measure the intensity of spontaneous facial actions.

2. *Deep-learning*: Feature learning approaches based on deep-learning methods are widely explored for HAR task because of their promising performance, robustness in extracting features and their generalization ability for different types of data. These methods are very data harvesting in the training process. They aim to learn multiple levels of representation and abstraction that allow a fully automated feature extraction process. Deep learning-based methods can be considered as trainable feature extractors, which allow

the recognition of high-level activities with complex structures. However, high computational complexity and huge data requirements for the training phase are still among ongoing challenging problems. Deep learning-based approaches can be categorized into:

(a) *Generative methods*: These are unsupervised models which are based on the famous quote of Richard Feynman: "What I cannot create, I do not understand." (212). They use unsupervised learning to represent any kind of unlabeled data distribution. The new representation reduces the data dimensionality and comply from the data distribution. Therefore, the main aim of the generative models is to understand the data distribution including the features that belong to each class, in order to replicate the initial true data distribution of the training set. The most commonly used and efficient approaches are: Auto-encoder (213, 214), Variational Autoencoders (VAE) (215) and Generative Adversarial Networks (GAN) (216). As an example of generative models, (217) proposes an end-to-end deep learning model for abnormal activity recognition in videos. The proposed architecture is similar to GAN, where the two networks compete to learn and collaborate in the detection task.

(b) *Discriminative methods*: these are supervised models that use a hierarchical learning model composed of different hidden layers to classify raw data input into various output categories. The most used are Deep Neural Networks (DNN), Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). As an example, in (191), the authors aim to demonstrate the advantages of long-term temporal convolutions and the importance of high-quality optical flow estimation for learning accurate video representations for human action recognition. For that, they investigate the learning of long-term video representations. They consider space-time convolutional neural networks and study architectures with Long-term Temporal Convolutions (LTC).

(c) *Hybrid models*: these methods integrate generative and discriminative models to gain performances and advantages of both models such as (214, 218).

3.3.2 HAR approaches according to the recognition stages

Human activities recognition systems are, in general, composed of three main steps: Detection which consists in determining the part of the body to follow or to recognize; Tracking that provides a connection between the successive images, and; Recognition which consists in interpreting the semantics of the localization, the posture, and the activity. In order to investigate further HAR methods, this subsection is devoted to represent different state of

the art vision-based HAR techniques associated to each stage. This assumes that the choice of the algorithm depends on the selected representation of activities. We classify methods for human activities recognition into the following:

1. *First stage methods (detection)*:

   (a) *Skin color*: This can be used to detect the desired body part. The problems of skin color-based methods reside mainly in choosing the relevant color space and with false detections because of the objects of the scene whose color is close to that of the skin. Very common solutions to these problems are the separation of brightness and chromaticity components, elimination of shadow effect and background subtraction (6, 219, 220). Other problems may influence the results of these methods, like the inherent discrepancy of skin color variations according to ethnicity group, lighting conditions and the sensitivity of the employed acquisition devices.

   (b) *Shape*: The contours of the body part shape can be extracted to follow and recognize human activities. To increase the reliability of these techniques, sophisticated approaches of post-processing are used for the extraction. Shape-based methods are typically independent of the camera view, skin color and conditions of lighting but present the difficulty of classifier's implementation. Moreover, background objects can be confusing or can occlude the forms to be detected. For instance, (78) uses the skeleton shape compared to information of the skeleton embarked in MICROSOFT SDK in order to detect human actions. In (8), the detection is based on appearance through a combination of the color and the shape to ensure a region segmentation that contains the person's hand. Shape-based methods were also used by (54, 221).

   (c) *Pixel values*: The appearance can be expressed in terms of pixel values change between images of a sequence according to the activity. Indeed, it is evident that the difference in appearance between various activities is more noticeable than among people performing the same activity. So, techniques based on the appearance of body parts are proposed as in (222).

   (d) *3D models*: They attempt to build matches between characteristics of the model based on various features of the images. These techniques are independent of the viewpoint, but present some limitations in terms of accurate positioning characteristics. For instance, (223) uses a 3D model method for hand gesture tracking.

(e) *Motion*: Motion can be detected using the difference in brightness of pixels of two successive images. For example, authors in (55, 224) use several motion characteristics based methods to detect body parts.

(f) *Anisotropic-diffusion*: This is based on the extension of the successful anisotropic diffusion based segmentation to the whole video sequence to improve the detection of object contours and regions of interest that may trigger specific activity, see (225, 226). Therefore such concept can be used to detect and describe activity patterns. For instance, the work in (227) uses anisotropic diffusion based approach to enable the recognition of crowd activities using a two-step clustering scheme that extracts the semantic regions and the coherent motion. In (226), an anisotropic filter is used for noise discrimination of 3D human activity recognition.

2. *Second stage methods (Tracking)*:

Various methods, independently from the ones used for the detection phase, can be used to track the human body. If the detection method is fast enough to operate at the frame acquisition rate, it may be the same in both stages. Tracking methods can also be categorized into the following:

(a) *Template-based methods*: these techniques use models to follow the body parts (221). They require continuous learning with high frame rate and are classified into two categories:

□ *Features tracking based on the correlation*: Regions of an image containing the body parts to be followed are used as a prototype to be detected in the following image. These techniques require that the part being tracked remains in the same neighbourhood in the successive images. They are influenced by the lighting variations and may not be efficient in real time systems. In (8), the authors use 3D information of depth maps to ensure the follow-up of the human body. Optical flow characteristics of two successive images are employed in (62) using the algorithm of Kanade Lucas Tomasi (228, 229). In (55), the hand gestures are followed using an articulated 3D model.

□ *Contours based tracking*: They are deformable contour techniques (snakes) which initially place a contour close to the area of interest, then it is warped in an iterative way using active shape models in each frame to make the snake converge (230). These techniques are more efficient if a contrast between the object to be tracked and the background exists. They can be used in real-time systems, may handle several targets at the same time and can also be adapted

to complex postures. On the other hand, they are influenced by color intensity variations and are limited regarding smoothing and softening of contours.

(b) *Optimal estimation*: These methods evaluate the state of moving systems from series of measures (231, 232). In the case of HAR systems, they are used to estimate the movements of the human body in similar way that Kalman Filter does (233). They can be used in real-time systems, and can deal with uncertainty, but have limitations against cluttered backgrounds.

(c) *Particle Filters*: These methods consist in following the positions of the body parts and their configuration in complex environments. A particular location of the part of interest is modelled by a set of particles.

(d) *Cam Shift*: They are based on the mean shift algorithm which use the models of appearance based on density to represent the targets (228, 229). Such methods consist in finding, in a sequence of images, the nearest model of distribution to the sample model using an iterative research. These methods are simple and have low-cost calculation. They have some limitations in complex scenes or scale variations.

3. *Third stage methods (classification)*:

(a) *Support Vector Machine (SVM)*: The classification is performed by the construction of a set of hyper planes in a multidimensional space separating the elements from various classes with a vast margin. SVM classifier is the most used hand-crafted classifier because it offers a very high rate of recognition performance and classification. It was used with a Gaussian, Radial basis function (RBF) or linear kernel with or without parameters in (8, 42, 45, 46, 53, 62).

(b) *Naive Bayesian classifier*: This is a probabilistic classifier based on the theorem of Bayes (234). It consists in counting the number of occurrences of the key motion in a video sequence. To recognize a new action, the rule of Bayes is applied and the selection of the activity which has the greatest probability a posteriori is done. In (62), a comparison between this classifier and SVM is carried out.

(c) *Algorithm of K_Nearest Neighbor*: it consists in the classification of objects based on the majority of their neighbors' vote (42, 46). To identify neighbours, objects are represented by position vectors in the multidimensional space of features. This algorithm is sensitive to local data structure. The work in (47) compares the obtained classification results using this algorithm and SVM.

(d) *K_means*: This is a clustering algorithm, which is sensitive to data structures and consists to iteratively calculate the k-distances from each class centroid to each datum (42). The points are then assigned to the nearest cluster and the centers are re-evaluated to be the average of their class values. The process is repeated until the stop condition is reached, which can be the maximum number of iterations or a tolerance level. It is used by (62) for the classification of human actions.

(e) *Mean shift clustering*: These are non-configurable techniques of clustering that do not require prior knowledge about the number of clusters and do not limit their forms.

(f) *Machines finite state*: In this model, the static gestures and postures are represented by states, authorized changes where temporal and/or probabilistic constraints are represented by transitions, and finally the dynamic gestures are represented by arcs between the initial and the final states. These machine-finite states require the modification of the model whenever a new gesture appears, and have a high computational complexity because they are proportional to the used gestures.

(g) *Hidden Markov Models*: These models represent solutions to the segmentation problem and are among the most popular (2, 43, 46). They constitute a generalization of Markov chains which are finite state automaton with a probability value on each arc. They have been discussed in (24).

(h) *Dynamic time warping*: This algorithm calculates the distances between each pair of possible points from two signals according to their associated characteristic values. This allows estimating and detecting the movement in a video sequence. It was used by authors of (53) in order to eliminate the problem of rate variations generated by the classification of human actions.

(i) *Neural networks*: They are mathematical models whose design is inspired from the functioning of biological neurons (43, 46). They are generally optimized by probabilistic learning techniques, in particular, Bayesian. Neural networks allow creating fast classifications which can be applied to real-time systems. For instance, (235, 236, 237) use neural networks for activities recognition.

3.3.3 HAR according to the source of the input data

Methods of human activities recognition are categorized, in (135), with regard to the nature of the detector into two main categories: Uni-modal methods that consider an activity as a set of visual characteristics and allow the recognition from single modality data (238, 239);

Multi-modal methods, which combine collected features from various sources, and therefore several modalities are used (61, 240). We can also mention the work of (45), where the authors combine speech and localization of human body for HAR.

Among the uni-modal methods, we can mention the space-time methods, which concatenate the time and the 3D representation of the body to locate activities in space, allowing a detailed analysis of human movements (194, 241). Often sensitive to noise and occlusion, these methods are not adapted to recognize complex actions, such as in (47, 62). On the other hand, the stochastic methods, which represent the activity by stochastic models like Markov Model, enable the modelling of human interactions and the recognition of complex activities. These methods yield approximate solutions whose training is difficult because of the large number of training parameters (176, 177). We can also mention rule-based methods that characterize the activity using a set of rules or attributes (43, 65). They present some difficulties during the generation of these rules and attributes, in the analysis of long video sequences and during the recognition of complex activities. Lastly, another family, shape-based methods has emerged. They use the shape features to represent and recognize activities (8, 55, 62, 242). The existence of a great number of pose estimation devices at low cost makes these techniques very useful, although they depend on the viewpoint, occlusion, people clothing and are sensitive to lighting variations.

The existing uni-modal vision-based approaches can again be categorized according to the input data type into many categories among which RGB images, skeleton data and depth images are the most commonly employed ones.

RGB images for HAR: Many works in the literature rely on the RGB videos to construct robust HAR systems. Several holistic action representations based on RGB images and powerful features are adopted by many authors to describe actions robustly. For instance, (243) combined shape features and optical flow calculated among RGB frames to detect change in motion. These approaches allow us to track the person in the scene and classify the ongoing activity. In addition, (190) proposed a long-term motion descriptor called sequential Deep Trajectory Descriptor (sDTD) which feeds a CNN-RNN network with dense trajectories to learn an effective representation for long-term motion. On the other hand, silhouettes were exploited by (162, 244) after extracting them using background subtraction. Automatic feature extraction from RGB images through deep learning was also suggested in many works. This is justified by the strong ability of CNN networks in dealing with RGB images. The authors of (245) extracted features from video sequences using pre-trained model, then an architecture of Deep Bidirectional LSTM for learning sequence information in the features of video frames was proposed. Similarly, (246) presented a Deep Long-term Recurrent Convolutional Network that deals with spatial and temporal features

at once. Deep learning methods represent low level to high level features with multiple layers of the neural networks. Activity classification is therefore performed either using a popular machine learning classifier or using deep learning networks. The above can also be combined as in (247) where a method that integrates graphical models and deep neural networks into a joint framework was presented.

Depth images for HAR: Due to the emergence of depth cameras that can overcome some inherent privacy and limitations issues related to traditional cameras, depth image-based representations for HAR have evolved significantly in recent years. 3D structure of the body can be generated by integrating depth sensors and body tracking, enabling straightforward action recognition. Many limitations related to lighting variations, perspective change, variation in appearance, complex backgrounds and scale variation can be resolved using depth-based HAR methods. Furthermore, the extraction of information content generated by depth images is often straightforward. This includes the body shape information, the silhouette data and the whole image region within camera view. Several depth-based HAR approaches have been suggested in the literature. For instance, a local spatio-temporal descriptor for action recognition from depth video sequences is developed by (248). It takes into account shape discrimination, motion change and action speed variations to distinguish between different actions. Similarly, (249) proposes a human pose representation model based on deep convolutional neural network CNN. The proposed model maps human poses acquired from several views of depth videos to a view-invariant high-level space. Although, depth image-based HAR has drawn growing interest by providing very promising results, depth-based methods still face difficult issues such as occlusion.

Skeleton-based action recognition: With the quick advent of depth sensors and algorithms of real-time skeleton estimation, many authors have demonstrated that skeleton features are more robust than RGB and depth features. This allows them to take advantage of this type of features for HAR. 3D locations and angles of joints are common features that can be used to build robust skeleton representations for HAR. Various methods based on skeleton analysis and representations of the set of joints for action recognition have been proposed in the literature. For instance, (250) proposes an end-to-end fully connected deep LSTM network for skeleton-based action recognition that relies on co-occurrence features of the skeleton joints. The co-occurrences of the joints are proven to be able to characterize accurately human actions. Similarly, (251) proposes a novel adaptive recurrent neural network (RNN) with LSTM architecture to automatically regulate observation viewpoints during the occurrence of an action. The 3D skeleton newly represented in a new coordinate system is used for accurate action recognition. Moreover, an end-to-end spatial and temporal attention model based on Recurrent Neural Networks (RNNs) and LSTM for HAR from discriminative

joints of the skeleton was proposed by (252). Likewise, authors of (253) presented an enhanced skeleton visualization method for view invariant HAR, where a sequence-based view invariant transform for the skeleton joints is performed and then the newly generated skeleton is visualized as series of enhanced RGB-images that encode spatial and temporal information related to skeleton joints. Finally, features were extracted using a CNN-based model allowing action classification.

Among multi-modal methods, there are emotional methods that associate visual and textual features to classify the emotional states of individuals in static images (7, 254). On the other hand, the behavioral methods aim to recognize the behavioral attributes like the emotions, mood, …etc. These methods allow the recognition of complex human activities using complex classification models, which make the specification of emotional attributes difficult (255, 256). Finally, we shall mention the methods based on social networks, which allow recognition of social events (wedding, birth …etc.), and interactions. They are limited by the number of people in interaction and the modelling difficulty in complex scenes (58, 257).

3.3.4 HAR approaches according to the machine learning supervision level

HAR is based on machine learning approaches and can be further categorized according to the level of learning supervision into three sub-classes; supervised, unsupervised and semi-supervised methods.

Supervised methods: In supervised learning, the training takes place offline and is performed by the machine using the well labeled data in order to predict outcomes of unforeseen data. For HAR, supervised methods are used to classify and recognize short term actions. Methods of this sub-class are computationally simple, highly accurate and trustworthy. We can quote as examples of supervised learning: Support vector machine, Linear and logistics regression, random forest and neural networks. Works of (13, 183, 211) are examples of supervised approaches for HAR.

Unsupervised methods: This refers to machine learning techniques that do not need any supervision mechanism. These techniques allow the model to discover by itself the discriminative features of the input unlabelled data using a real-time learning such as K-means and K-nearest-neighbor. For HAR, unsupervised methods perform well for finding spatio-temporal patterns of motion and generating scene models in order to automatically localize activities. These methods are computationally complex, less accurate and trustworthy. For instance, (258) presents a new algorithm that models the human activities in a completely unsupervised setting enabling to recognize daily living and forgotten activities. The probabilistic model considers, both the short-range and the long-range action relations and shows considerable results in action segmentation and clustering. Similarly, unsupervised approaches are

presented in (201, 214, 239). Authors of (214) present a multi-layer Long Short Term Memory (LSTM) networks to learn representations of video sequences, while an auto-annotation framework to iteratively label pedestrian instances using multi-modal data is introduced in (201).

Semi-supervised methods: This refers to hybrid methods which combine supervised and unsupervised learning. The training is performed using both labeled and unlabeled data, mostly when the labeled data is not enough to product accurate model. For HAR, these methods can benefit from discriminative power of supervised approaches to distinguish between features and the ability of unsupervised methods to automatically localize actions. For instance, (59) presents a hybrid framework for online recognition of daily living activities. The authors provide a complete presentation of human activities by exploiting both unsupervised (to represent global motion patterns and localize activities) and supervised learning approaches (to distinguish between actions occurring under specific scene region). Other works such as (65, 163, 181) provide also semi or weakly supervised framework for HAR in videos.

3.4 Limitations

This section is a presentation of various issues that may affect the effectiveness of HAR systems. Some of them are specific to the methods used during the various phases of the recognition process. Others are related to the acquisition devices, experimentation environments, or various applications of these systems. Lighting variation is the main difficulty facing vision-based recognition systems in general, because it affects the quality of images and thus, analyzed information. In the same way, perspective change is another limitation of the current systems which were implemented to operate using a single view acquisition devices. This problem reduces the amount of extracted information and provides a limited visualization of activities being analyzed. This includes occlusion with its different types: self occlusion where body parts occlude each other, occlusion of another object and partial occlusion of human body parts are major limitations to HAR systems. These problems are discussed in the works of (3, 10, 47, 53, 55, 59). The variety of gestures linked to the complex structure of human activities and the similarity between classes of different actions can also be source of additional difficulties due to data association problem it may involve. This needs to be addressed in order to set up complete, robust and flexible HAR systems under various conditions or environments. The limitation of hand configurations, predefined actions or postures, and recognition of simple gestures or activities are discussed in (3, 8, 10, 45, 55). Some methods of detection based on the form or the appearance, such as colorimetric segmentation, can confuse the human body parts with objects of the scene, as in (8, 10, 42, 55)

or cannot operate correctly during variation in appearance or clothing of people such as in (3, 47). These problems are associated with other problems related to methods of recognition or acquisition devices, such as noise (55), complex or moving backgrounds and unstructured scenes as in (3, 10, 42, 47, 53, 59), and scale variation when the person gets closer or move away from the camera (3, 47). Finally, many researchers rely on their own recorded datasets to test the performance of their proposals. This raises the concerns about the limitation of available benchmark dataset for testing novel domain-specific applications. We can mention for instance the case of daily life activities and fall detection datasets which are not large enough to be used for training effective models.

3.5 Challenges of the recognition systems

The current HAR systems present a big number of challenges they have to cope with in order to provide the principal functions for which they are developed. For instance, most HCI or video surveillance systems based on HAR must provide continuous monitoring and generate reliable answers at the right time in order to ensure the performance of the provided results. This challenge is discussed in (8, 10, 55). Furthermore, this becomes a major issue when modelling and analyzing interactions between people and objects with an appropriate level of accuracy which is still challenging. This would be very useful for surveillance and public security applications and may help to detect several abnormality scenarios.

The use of these systems for the aim of surveillance, elderly assistance and patient monitoring together with the increasing implementation costs raise also new societal challenges: acceptance by the society, privacy, side effects of installation of these devices at home as well as large scale applications. For instance, authors in (45, 55) discuss the challenge of device integration at home for tracking, which is considered as violation of intimacy and privacy. To address this last problem, it is interesting to explore the development of HAR systems on smartphones. This may contribute to overcome the user's privacy constraint since the recorded data would be stored on his own device and may reduce also the computational time related to the transmission between the distant server and the device. However, on-device implementation is a challenging issue as well due to memory constraint, high number of parameters needed by the recognition model and the battery life of the device (135). Another challenge is related to implicit dependency of such systems with the physical and physiological abilities of the user. Intuitively, HAR systems should not depend on the user's age, color, size or capacity to use such systems. Both experienced user and beginner should be able to use these systems, and in the same way. This challenge is raised in (10, 55). The gestures independence and gestures spotting from continuous data streams constitute also another type of challenge, since it is still difficult to localize temporally the gesture in long

continuous videos. Indeed, HAR systems are not yet able to detect and recognize various gestures under different background conditions and are not tolerant with the scalability and growth of gestures. This is the major challenge which is studied and considered by many works such as: (3, 8, 10, 47, 55). One more area that need to be explored is the ability of HAR systems to be context-aware. This may help to make use of proposed approaches and progress made in many application domains.

Understanding and detection of daily life activities in long-term videos is a challenging task. This is due to the fact that the long-term videos containing daily life activities are composed of several complex activities. These activities are difficult to model because of their complex structure as well as the big variation in ways of performing the same activity. Another issue is related to the overlapping between starting and ending time of each particular activity. This challenge is addressed by (59). In addition, the discrimination between intentional and involuntary actions is still very challenging area to tackle.

Other challenges are discussed in (2, 46), which are human activities recognition through missed parts of video, recognition of more than one activity performed by one person at the same time, and early recognition and prediction of actions, especially in crowded environments. Nowadays, memory constraint, high number of parameters update, collection and fusion of large multi-modal variant data for the training process as well as deployment of different architectures of deep-learning based methods in smartphones or wearable devices are still unresolved issues in deep-learning HAR systems (135).

3.6 Conclusion

The surveys discussed above do not cover all aspects of human activity recognition. They highlight different taxonomies, applications and specific theoretical angle of HAR. This motivates us to introduce this deeper analysis of human activity recognition, although our review here pays a special attention to vision-based HAR systems. We analyze approaches proposed in the literature according to different criteria. We started by presenting HAR techniques using handcrafted and learning based features for the features extraction process. Then, we enumerate different techniques used for the three stages of recognition consisting in detection, tracking and recognition. We discussed also HAR techniques using only one modality versus multiple modalities. We consider as well supervised, semi-supervised and unsupervised techniques. In a second time, we discuss the limitations and the challenges we have to cope with to implement robust and efficient HAR systems.

4

Overview of Fall Detection

4.1 Introduction

Performing regular daily life activities by the elderly population can be affected by many serious health issues among which fall and its resulting injuries are the most frequent (259). This is mostly experienced by the elderly because the brain and nervous system go through natural changes (260). Indeed, nerve cells and weight across the brain and spinal cord are lost and cannot be regenerated. So, plaques and tangles are formed by collecting tissue of waste products in the brain, which can cause abnormal changes. Therefore, nerves' messages may not be delivered, resulting in problems with movement and safety issues. Moreover, falling is due to other inherent factors such as age-related biological changes, neurological disorders, physiological health profile and environmental conditions (260). The authors in (260) presented a detailed study of the different factors that may lead to falls in elderly population. In general, falls can result from sudden loss of balance, stability, dizziness, or vertigo during daily life movements. It can also be caused by chronic diseases, cognitive impairment, using a walking aid or multiple medications, gait and visual deficit (261). Falling is a major health problem in elderly population that has recently attracted the attention of many researchers. One can define fall as unintentional or sudden change of position of the body from an upright, sitting or lying to a lower inclining position (31).

In this chapter, we explore some fall detection related concepts that help us to introduce the FD field. So, we introduce some important definitions in section 4.2 such as the meaning of fall, surveillance and monitoring and smart homes. Then, we present in section 4.3, the three types of fall events. Section 4.4 is devoted to provide the main fall detection application fields, especially in smart homes and elderly monitoring. Afterwards, we summarize some existing Fall Detection approaches in section 4.5, we enumerate some benchmark datasets used to validate those approaches in section 4.6 and discuss FD related limitations in section 4.7. Finally, we summarize the chapter in section 4.8.

4.2 Definitions

In this section, we introduce some definitions of main concepts related to Fall detection. These definitions may help to introduce the studied field.

4.2.1 Fall

Fall is defined in (262) as an event which results in a person coming to rest inadvertently on the ground or the floor or any other lower level. A fall event can be decomposed into pre-fall phase, critical phase and post-fall phase and recovery. The pre-fall phase consists in the moment before the occurrence of the fall which is usually the reason behind the fall. The critical phase corresponds to the moment when the fall occurs due to fainting, tripping or slipping. Finally, the post-fall phase and recovery consists in the after fall phase where the person may still lying on the floor or has woke up.

4.2.2 Fall Detection

Fall detection consists in recognizing a fall among all daily life activities.

4.2.3 Surveillance and Monitoring

Surveillance refers to the act of close observation of an individual, a group of people or places by visual, electronic, photographic or other means. From the other hand, monitoring consists in listening, observing and recording data to maintain or improve security and quality of life of the monitored individual.

4.2.4 Smart Home

Smart home (or home automation) refers to a residence equipped with different types of sensors and internet-connected devices that can be remotely monitored (263). This technology enables to respond to the inhabitants needs and requirements by controlling different factors such as temperature, air conditioner, lighting and security access to the home. Besides, smart home technology contributes to health and well-being enhancement by allowing recognition of health status, analysis and prediction of the resident actions to promote his convenience, security and comfort.

4.2.5 Activities of Daily Living

The term of Activities of Daily Living (ADLs) is mainly used in healthcare to describe fundamental skills that people carry out independently without the need of others help. It is used to indicate the person's functional status and its ability to be autonomous (264).

4.2.6 Ambient Assisted Living

Ambient Assisted Living (AAL) consists in using information and communication technologies and smart devices in a person's home and work environments to make his life easier. AAL is based on HAR and it allows to ensure the elderly safety, extend the time of his independent living and monitor his healthcare (263).

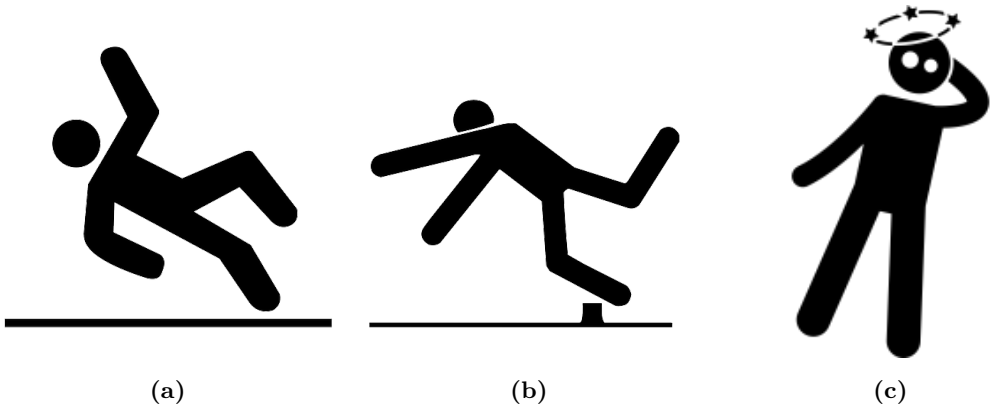
4.3 Fall Types

Falling is an abnormal human activity that may occur in many different ways. The impacts and consequences can vary drastically depending upon the fall type because, for example falling whilst standing could be more harmful than whilst sitting on a chair. Roughly speaking, different kinematics characteristics during falls could be observed. For instance, many studies exploited the inertial velocity profile of the body such as (265, 266, 267). So, to explain the aforementioned example, lets bring up the vertical velocity of landing from a standing fall which is larger than the one of landing from a sitting fall and is larger again than the velocity in normal activities. Another important aspect is the direction one takes whilst falling, which also depends on the fall type. For instance, faints and slips resulting in sideways and backwards falls may absorb the kinetic energy of the fall since they can lead to landing first on the knees or the hands which may reduce the impact of the fall. Fall types are categorized according to (268, 269) into three main types as follows (See Figure 4.1):

- Backward fall: refers to falling to the back with impact on the hip or buttocks. This could be caused for example due to a backward disequilibrium or a slippery floor and it seems to have the lowest rate of avoidance and the highest rate of full body impact (270).
- Forward fall: refers to falling flat on one's face and is mainly caused by tripping. This kind of falls has frontal impacts and occurs mostly among adults and elderly.
- Side-way or lateral fall: refers to falling on one side due to fainting and has impacts mainly near the hip of the body.

4.4 Fall Detection Applications

Fall detection is a very important research topic that has been largely studied in the literature. FD systems were mainly used in smart homes for home security, healthcare and home automation purposes aiming at enabling elderly isolated population to live alone for as long as possible.



(a) (b) (c)

Figure 4.1: Fall Types: a) Backward fall, b) Forward fall and c) Lateral fall due to fainting.

Adults, especially the elderly, are spending the majority of their time in their home or workplaces where remote assistance should be envisaged. This motivates the construction of smart homes and home automation. Integrating fall detection system in a home automation process is essential, since it helps the identification of falls before their occurrence, alerts caregivers and yields necessary assistance. Providing enough support to the elderly at required time reduces their dependency to a great extent and hence decreases the number of required caregivers. FD systems evolve as well in the automation of homes to maximize the user comfort, safety and luxury.

In general, observing and monitoring the cognitive and physical capabilities of the inhabitants and patients among time could be very helpful for doctors. Actually, detecting any changes may conduct to identify emerging medical conditions and anomalies before they become critical.

Moreover, occupants may feel safe while being watched and may benefit from efficient care whenever needed. Indeed, once falls or abnormal activities are detected, the system is able to alert the caregivers who can respond urgently. This can by far enhance the home security of the individual.

4.5 Fall Detection Approaches

Fall detection techniques can be categorized into three major classes: ambient-based, wearable-based and vision-based systems (261). Ambient-based systems use light, proximity, motion, and vibration sensors to collect daily life activities data and detect falls. Wearable-based systems rely on the sensors embedded in particular devices that the subject should wear in order to track his/her motion (271). Additionally, vision-based systems use RGB or depth

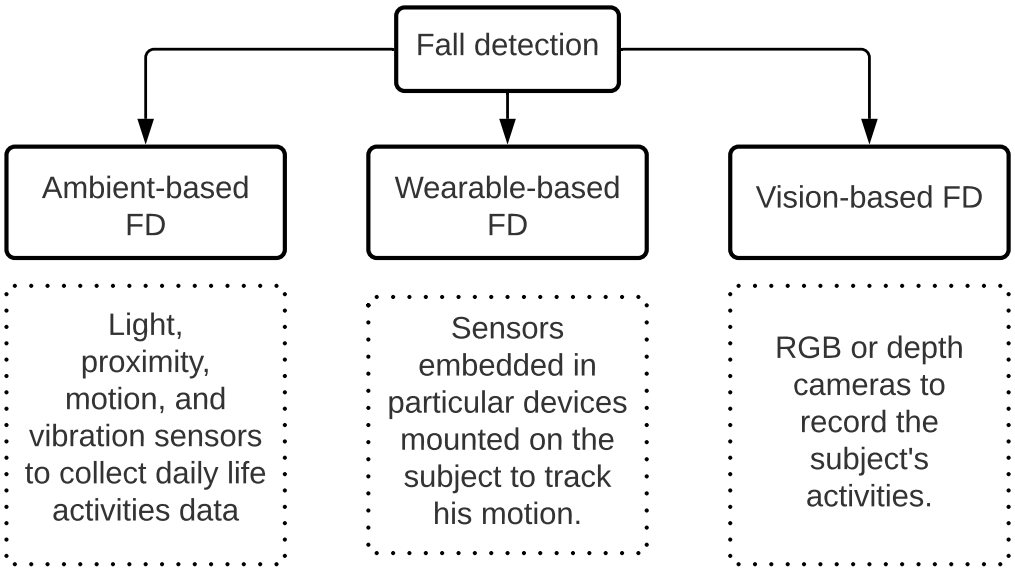


Figure 4.2: Categorization of Fall Detection Approaches

cameras to record the subject's activities, in indoor or outdoor environments (271). Figure 4.2 summarizes the existing fall detection approaches. The recorded images or videos are analyzed later to detect falls. In this thesis, we focus on the vision-based FD techniques.

Roughly speaking, Vision-based FD approaches focus on meaningful fall related features extracted from the video frames such as silhouettes, body shape and skeleton information. These features are then used as input to some machine learning classifier such as SVM, KNN, Hidden Markov Models (HMM), among others, to train and later automatically detect fall and non-fall cases. For instance, (272) extracts distinctive features of human silhouettes to construct new action representations. The authors model the actions using a bag-of-words (BOW) and conduct the classification using an extreme learning machine (ELM). Authors in (273) suggest robust features called History Triple Features using a generalization of the Radon Transform. Furthermore, SVM based approaches have proven their efficiency for fall detection tasks in many alternative works see, for instance, (274, 275, 276). In (274), five distinct features are employed (aspect ratio, change in aspect ratio, fall angle, center speed and head speed). Authors in (275) use a normalized motion energy image (MEI) to model the silhouette shape deformation features, while (276) proposes a novel descriptor, called Trajectory Snippet Histograms, to model the rapid motions change. They used BOW to describe each video clip and train an SVM for unusual videos classification. In addition, shape and motion features are tracked to detect falls using a single camera based system in

(277). Likewise, (278) proposes to analyze dynamic appearance, shape and motion features of the target person and then characterize the human falls with simple velocity statistics of moving features. Authors in (279) suggest a vision-based fall detection system for elderly living alone. The system relies on the optical flow estimation to estimate the speed of motion and to deduce the fall activity accordingly, while comparing the last positions of the target. Deformation of the body shape, the centroid, the perimeter and the principal axis of the silhouette are significative features for fall detection. This is due to the fact that changes may happen drastically and rapidly during falls than during normal activities. Based on this, The authors of (280, 281) proposed to detect falls using silhouette deformation analysis techniques during and after the fall. In addition, the authors of (282) proposed a video-based fall detection system based on multiple shape and motion features where the head and the vertical velocity of the head are determined to recognize falls. Moreover, a novel vision–based fall detection approach characterized by the utilization of human postures was provided by (283) where a fall-like accident is detected by counting the occurrences of lying postures and the fall is then determined by immobility verification. The system is composed of three major steps: human body extraction, human posture description and fall event recognition.

On the other hand, many vision-based research is devoted to fall detection using Kinect sensors. This is because depth cameras can overcome some privacy issues related to traditional camera systems. For instance, (284) proposes a real-time fall detection system based on 3D Kinect depth maps. These depth maps are used to extract 3D silhouettes features. Similarly, (285) employs Kinect sensor to acquire point cloud images and extract energy fall features. Other researchers demonstrate that using Kinect sensor alone does not provide sufficient coverage and, therefore, cannot yield robust and efficient fall detection capabilities.

With the advance in Deep Learning (DL) approaches, many researchers put forward DL based approach for fall detection tasks. For instance, (286) proposes a real-time fall detection approach that allows the capture of RGB video streams, individual's position estimation and, thereby, fall detection likelihood, which then generates potential alert messages to caregivers with registered audio and video. In (287), the authors present a novel FD method based on Convolutional Neural Networks (CNN) using optical flow images. Moreover, transfer learning is widely used to take advantage of pre-trained models by reusing their network weights or fine-tuning the classification layers. For instance, (288) was able to efficiently detect falls using a CNN Alexnet architecture. In (289), the authors present a two-stream approach based on MobileVGG network. Similarly, the authors of (289) combine an improved lightweight VGG network and the motion characteristics of the human body. Likewise, a 3D CNN-based method combined with long short-term memory (LSTM) is also presented in (290). The 3D CNN is used to extract motion and spatial features while the LSTM-based

spatial visual attention scheme is incorporated to locate the fall in each frame. Authors in (291) presented a fall detection system based on LSTM, using location features from the group of available joints in the human body.

4.6 Fall Detection Benchmark Datasets

To validate FD techniques, researchers need to use public datasets that are specific to the FD field. This allows them to get benefit of the particular characteristics that distinguish falls from other activities. In fact, there exist few publicly available FD benchmark datasets and all of them contain simulated data that was recorded by young healthy volunteers. We classify FD datasets according to the sensing technology been used to record data into contact-based, vision-based and multi-modal datasets as suggested by (292).

1. Contact-based FD datasets:

    - DLR dataset (293) was collected in 2010 by the Institute of Communications and Navigation of the German Aerospace Center (DLR) from one inertial measurement unit placed on the belt. Data of 16 adults performing seven activities (walking, running, standing, sitting, laying, falling and jumping) was recorded and manually annotated by an observer. In total, the dataset contains 4.5 hours of labeled falls and daily life activities.
    - MobiFall fall detection dataset (294) was implemented by the Biomedical Informatics and eHealth Laboratory of Technological Educational Institute of Crete. It contains data of 11 adults who performed 4 types of falls and 9 ADLs, with 6 trials for each activity. The data was recorded using accelerometer and gyroscope sensors of a smartphone positioned in trousers pocket. ADLs were chosen based on their commonness and similarity to actual falls. Figure. 4.3 illustrates a simulated fall from the MobiFall FD dataset.
    - The tFall dataset (295) was developed by EduQTech (Education, Quality and Technology) group of the University of Zaragoza. It contains data of 10 adults simulating 8 types of falls, repeated 3 times per subject, and ADLs recorded using two smartphones in their pockets.
    - Vilarinho et al.(296) created a dataset recorded using a smartphone carried in the thigh pocket and a smartwatch worn on the wrist of the subjects. The dataset contains 12 tyes of falls and 7 ADLs carried out by 3 subjects.

Figure 4.3: Simulated side-way fall from the MobiFall FD dataset (294)

- UMAFall (297) contains data of 3 types of falls and 8 ADLs, realized by 17 subjects recorded in domestic environment. The data was captured using a smartphone worn in right thigh pocket and four wearable sensors worn in ankle, waist, right wrist and chest as illustrated in Figure. 4.4. The subjects replicated every action at least 3 times.
- SisFall (298) contains 15 types of falls and 19 ADLs of 38 participants composed of 15 elderly and 23 young adults. Data was recorded using a self-developed embedded device with two accelerometers and one gyroscope. This device was fixed to the waist of the participants as illustrated in Figure. 4.5.

2. Vision-based FD datasets:

- SDUFall (272) is a depth action dataset that was built using one Kinect camera. It contains 6 types of actions (5 ADLs and 1 fall) performed by 10 young volunteers. Each subject carry out the action 30 times, where different environment conditions are randomly changed at each trial. The dataset has a total of 1800 video clips and some examples are shown in Figure. 4.6.
- The Le2i FD dataset (300) was collected using a single fixed camera in four different locations. It contains 221 RGB videos of 131 falls and 90 ADLs simulated by several actors. The dataset illustrates many difficulties of realistic video sequences

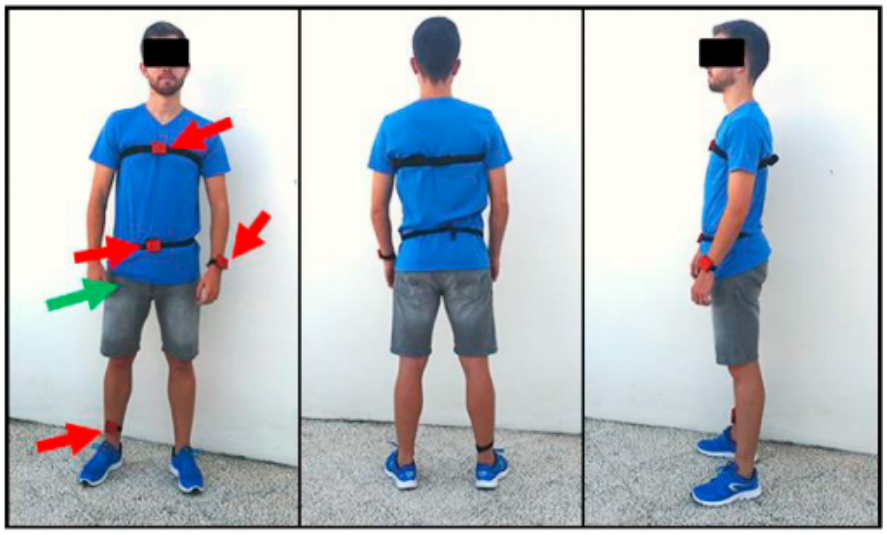


Figure 4.4: Location of the sensors (red arrows) and the smartphone (green arrow) on the subject for the UMAFall dataset (297)

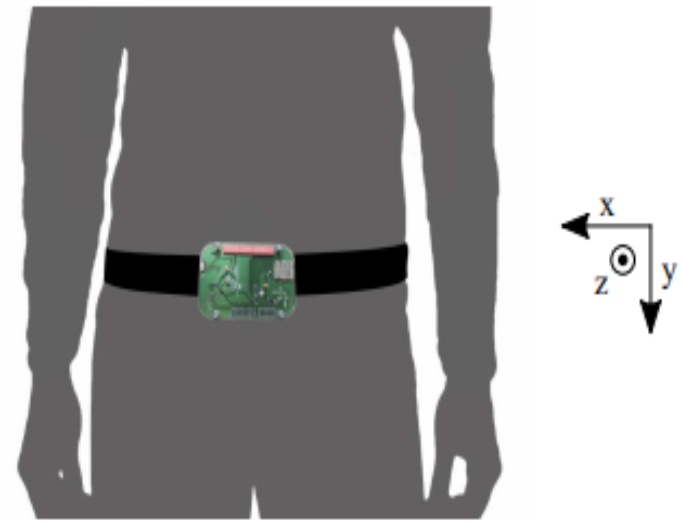


Figure 4.5: Location of the self-developed device used for acquisition for the SisFall dataset (298).

Figure 4.6: Fall alarms on sequence of images from the SDU Fall database (299)

of an elderly home or office such as variable illumination and occlusion. Figures. 4.7 and 4.8 illustrate some examples of ADLs and falls (respectively) from the Le2i FD dataset.

- EDF FD dataset (301) was captured in a non-occlusion settings, using two Kinect cameras fixed at two different viewpoints, where same event is recorded from both viewpoints at the same time. Data of 5 volunteers performing 2 falls, each with 8 directions in each of the two viewpoints was recorded with 3 ADLs that are similar to falling. The dataset contains in total 160 falls and 30 ADLs.
- OCCU FD dataset (301) was recorded in the same settings as the EDF dataset except the non-occluded environment and that each viewpoint was recorded at separate times from the other viewpoint. Similarly, 5 subjects performed 6 occluded falls in two viewpoints. The dataset contains 60 falls and 80 ADLs. Examples of occluded falls are illustrated in Figure. 4.9.
- Mastorakis et al. (284) created a FD dataset that was captured using a Kinect sensor attached to a tripod at a height of 204cm, inclined to the floor plane. The dataset contains 184 video samples of 48 falls and 112 ADLs simulated by 8 subjects. Samples from this dataset are given in Figure. 4.10.

3. Multimodal FD datasets:

- The UR FD dataset (302) was developed by the university of Rzeszow, using 2 Microsoft Kinect cameras for the fall events and only one camera for the ADLs.

Figure 4.7: Examples of ADLs from the Le2i FD dataset (300)

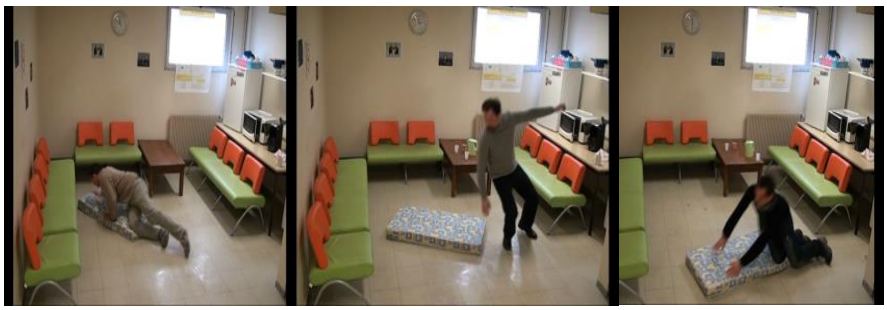


Figure 4.8: Examples of falls from the Le2i FD dataset (300)

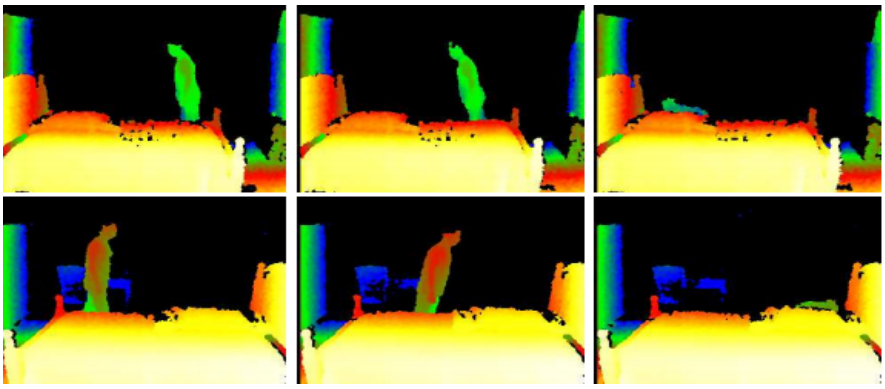


Figure 4.9: Examples from the OCCU dataset representing an occluded fall. The top row shows an occluded fall in the first viewpoint while the bottom row shows an occluded fall in the second viewpoint (301)

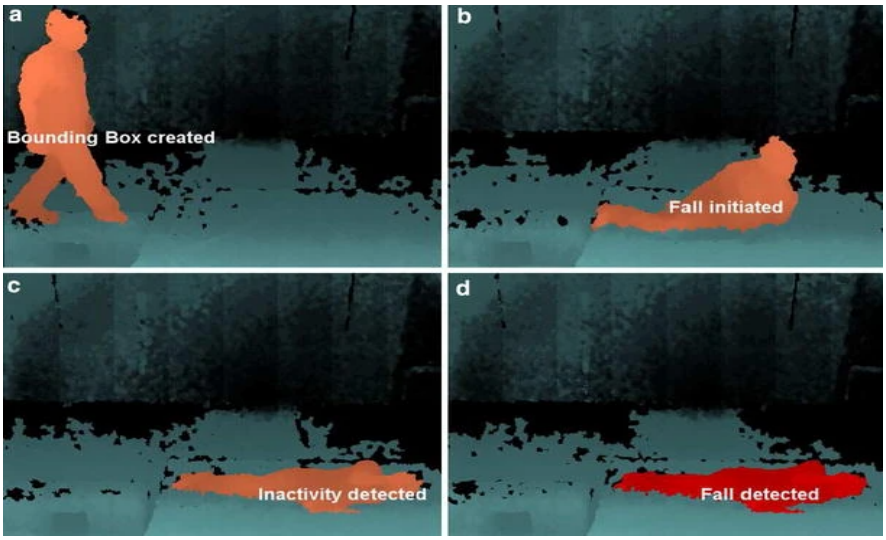


Figure 4.10: Examples of a sideways fall from the FD dataset created by Mastorakis et al.(284).

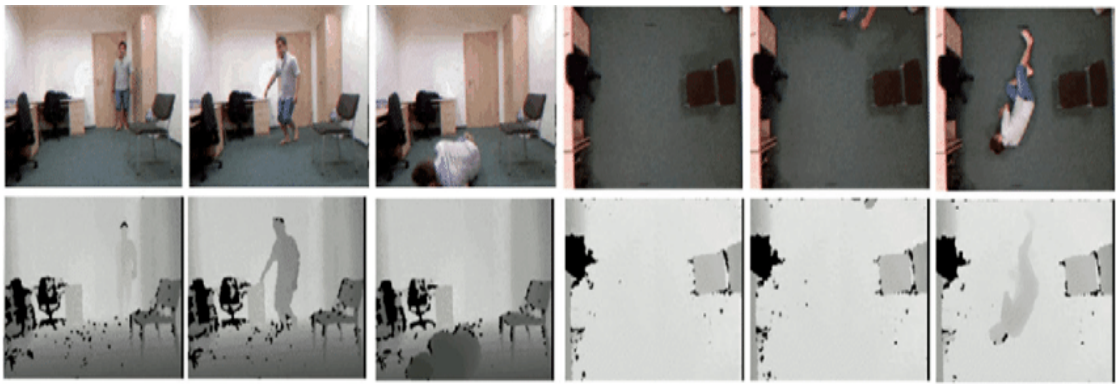


Figure 4.11: Samples from the UR FD dataset where the first row contains RGB images and the second row the depth images (302)

The dataset contains 70 (30 falls + 40 activities of daily living) sequences performed by 5 volunteers. Inline with RGB data, the dataset contains depth images and data collected from an IMU inertial device connected via Bluetooth. Some samples from the UR FD dataset are shown in Figure. 4.11.

- UP FD dataset (292) is a large multi-modal dataset captured at the Faculty of Engineering, Universidad Panamericana, Mexico. It includes 11 activities composed of 6 ADLs and 5 types of falls, performed by 17 young healthy subjects 3 times per activity.
- MultiCam Dataset (303) was captured using multiple IP video cameras (8 cameras). It contains 24 scenarios simulated by a healthy adult. The dataset includes

Figure 4.12: Samples from the multicam dataset demonstrating falling events (303)

22 scenarios of falls and confounding events and 2 scenarios of only confounding events. Figure. 4.12 illustrates some samples of falls from the multicam dataset.

4.7 Fall Detection Limitations

Although the considerable progress made in FD, current systems still suffer from diverse limitations that alter their robustness and reliability. In general, many HAR techniques were adapted to detect and recognize falls among daily life activities. However, FD should be considered as a separate research axis and proposed methods should deal particularly with this kind of activity and its specific characteristics. Furthermore, types of falls are not taken into consideration in most of the state-of-the-art works. All falls are studied in the same manner and solutions are proposed regardless the particular characteristics of each fall type. Besides, the existing FD benchmarks used in the literature are recorded by young healthy volunteers and contain simulated fall events that do not cover all fall types and do not yield to generalized models. Also, falls are relatively rare compared to other activities. This makes their occurrence in real settings uncommon conducting therefore to unbalanced data.

Limitations and challenges of HAR systems discussed in the previous chapter are as well applicable to FD systems. A particular focus could be given to occlusion. Indeed, the monitoring system is mounted in most cases indoor and it should be able to ensure the continuous follow-up of people over time. It should also deal correctly with surrounding objects or other individuals in the monitored environment. Again, privacy concerns should be addressed to make users feel safe, comfortable and independent when being watched.

4.8 Conclusion

In this chapter, we defined main concepts related to FD field to introduce the field to the reader. Then, we enumerated the different fall types with their underlying causes. We analyzed afterwards FD techniques proposed in the literature, which we categorized into three main classes: ambient-based, wearable-based and vision-based FD. Next, we highlighted some of the existing FD benchmark datasets that have been used in prior works. Finally, we discussed the limitations encountered by the FD systems. In the next chapters, we outline our proposed methodologies for HAR and FD. It is to note that even if lot of research has been made so far for FD, existing systems are still limited and inefficient. Therefore, a good FD system should be able to ensure the safety and the privacy of the elderly by enabling the quickness of intervention when a fall occurs.

5

Multi-Modal Vision-Based Human Activity Recognition using deep learning

5.1 Introduction

The increasing progress in sensing technologies prompted the emergence of intelligent real-time systems that can potentially impact the development of efficient human activity recognition systems and enhance the quality of life and security of the individuals (304). Major vision-based HAR works focus on using one single sensor modality to classify activities. This yields some limitations while discriminating complex activities due to environment conditions such as lighting, perspective changes and occlusions (249). To achieve good results and enable robust HAR systems, it is important to exploit more than one modality and to this end, different fusion strategies are to be explored (135). In this respect, we propose to combine three modalities from RGB, depth and skeleton data for vision-based HAR in order to achieve high recognition accuracy. Our framework integrates three sets of images created using data acquired from the modalities mentioned above. For RGB and depth data, we use an approximated version of rank-pooling in the same spirit as (33, 34) to create two sets of dynamic images. Each dynamic image summarizes information contained in the video frames into one single visual image. Furthermore, a skeleton data is used to create images that encode the locations of the skeleton joints among the video frames and hence describe the temporal aspect of the action. These newly created images are then employed to enable transfer learning from a pre-trained model in order to extract significant features. A feature-fusion strategy is performed later on using the Canonical Correlation Analysis CCA (305) to create highly discriminative feature vector that combines selective features from the

three single feature vectors. Once the underlined unified vector is created, we train a Long Short-Term Memory LSTM network to classify activities from the video sequences.

In this chapter, we present a new framework for multi-modal vision-based HAR. We introduce in section 5.2 the multi-modal human activity recognition field, where we explore some of the existing methods. We also highlight the video representation and the rank pooling techniques in section 5.2.1 and 5.2.2. Then, section 5.3 is devoted to outline our proposed framework where we give details of our proposal in different steps. Following this, we discuss the experimental setup and the obtained results in section 5.4, where different variants of our proposal are evaluated on publicly available UTD-MHAD and NTU RGB+D datasets. Finally, we conclude the chapter in section 5.5.

5.2 Multi-Modal Human Activity Recognition

Multi-modal data fusion in HAR consists in combining many sensor modalities data in order to increase the robustness and the reliability of the recognition system while reducing single sensor effects such as noise (135). To achieve this, it is essential to provide a complementary highly discriminative fusion of these modalities. In the literature, many fusion strategies have been employed to efficiently select meaningful information among different combined modalities (135). Feature-level fusion, through increasing feature-space, projecting on some external frame, or using correlation-like analysis, is one of the best strategies for fusing heterogeneous modalities (135). For instance, depth data, skeleton information and RGB images provide important complementary features. Indeed, depth data is more robust to illumination changes and scale variation but sensitive to occlusion; while skeleton information is more robust to occlusion effects and RGB image provides fine-grained image segmentation. Many vision-based HAR approaches combine two of these three modalities to improve the recognition accuracy but very few works focused on the combination of all of the three modalities. For instance, a robust HAR approach combining skeleton and RGB data streams was presented by (304), although, the authors used decision-level fusion instead of feature-level fusion. To effectively fuse features extracted from several modalities, some works use Canonical Correlation Analysis which allows them to learn from heterogeneous data and afford high linear correlation outputs. For instance, (306) developed a deep canonical correlated analysis to fuse accelerometer and gyroscope data for human activity recognition. We presented in Chapter 3, a categorization of human activity recognition approaches according to the source of the input data, where we discussed uni-modal and multi-modal HAR.

5.2.1 Video representations:

On the other hand, the temporal dimension of the action is often taken into account explicitly to enhance the recognition performance. However, many approaches extract spatial features of the image and deal with the temporal variations in the classification stage. In line with our overview on multi-modal HAR, we review related work in video representation.

Features extracted from image sequences expand to variations in action execution, person appearance, shape and motion. These should be sufficiently distinctive to allow distinguishing between different actions. Efficient action representation is the key to yield robust and expressive features. Therefore, many video-based HAR methods are based on video representation to efficiently describe the action. They can be grouped accordingly into two main categories. The first one considers video as a stream of still images or as transitions between frames. The second category represents videos as 3D dimensional volumes. The majority of hand-crafted-based or deep learning-based human activity representations belongs to the first category such as (34, 307). The popularity of the first category raises from its efficiency and simplicity of use for activity recognition. Moreover, video is represented as a spatial-temporal volume by stacking frames over a given sequence and action recognition is performed based on either spatial or temporal features or both. These features may be texture, color, posture, histograms of optical flow or histograms of oriented gradients. Many authors use spatio-temporal templates and 3D CNNs to learn features from spatio-temporal volumes and capture dynamics. For instance, (308) uses spatio-temporal templates, while (309) uses 3D CNNs for activity recognition based on video volume representations. Furthermore, a multi-view system to understand, in real time, the interactions between the ball and the players based on their respective 3D trajectories was presented by (310).

5.2.2 Rank pooling videos:

Rank pooling in videos allows to capture the video-wide temporal evolution while preserving actions execution temporal ordering. It was introduced by (34, 307). The authors of (34) proposed to train a linear ranking machine on the video frames and to use its parameters as a new video representation. When trained on different samples of the same action, the authors demonstrated that the ranking machines would have similar ranking functions. (311) extended the rank pooling to encode video sequences at multiple levels recursively where the output of each encoding level is itself the input of the next encoding level in order to capture higher-order dynamics. Similarly, (33) introduced a CNN-based approximated rank pooling approach that allows us to learn dynamic image networks for action recognition.

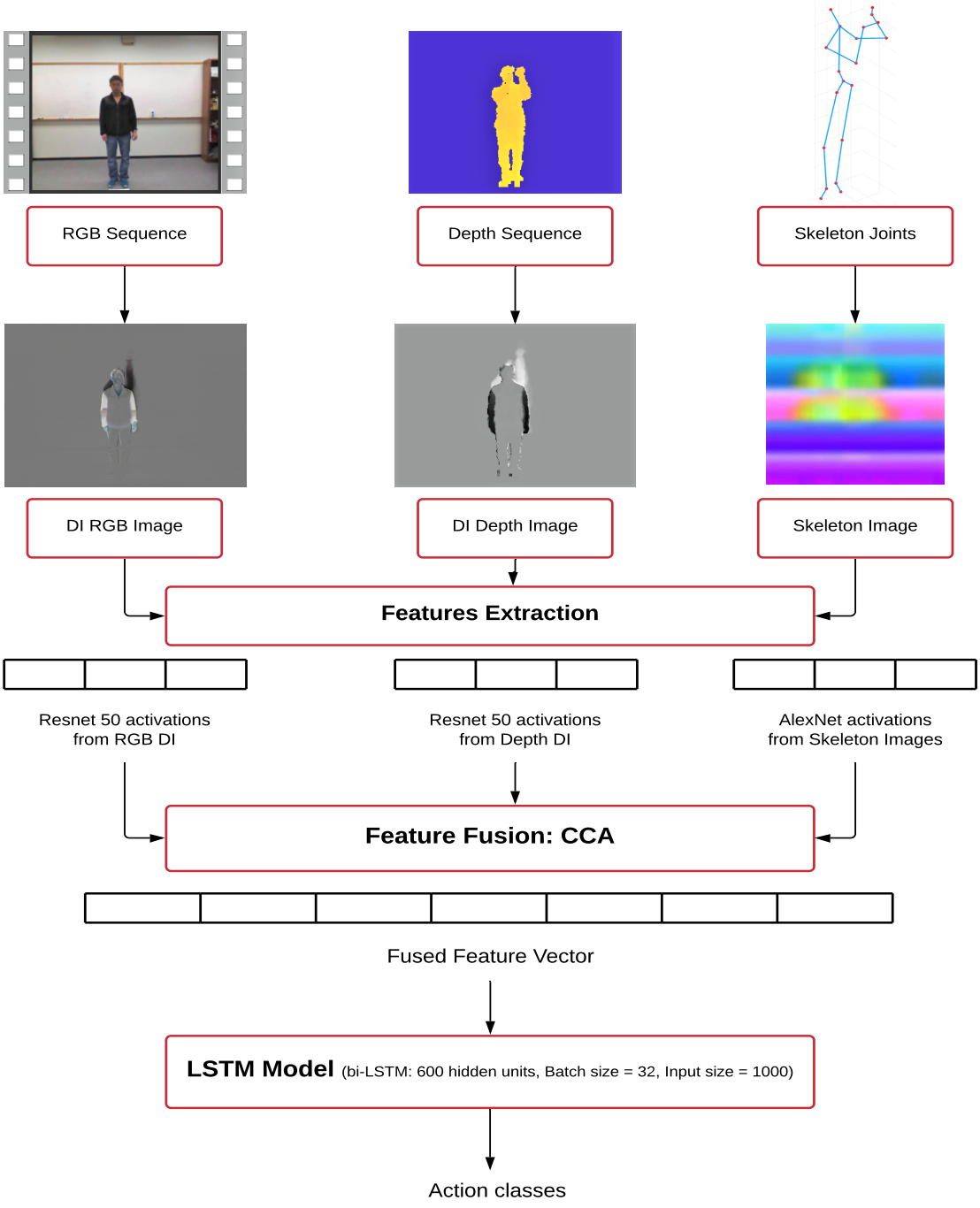


Figure 5.1: The general overview of our proposed vision-based multi-modal approach for HAR

Building on (306), our work combines features extracted from dynamic images and skeleton images using canonical correlation analysis. Dynamic images were calculated from RGB and depth sensors separately, while skeleton images refer to RGB image representation that we derive from skeleton joints information.

5.3 Our proposed methodology

We advocate a feature-level fusion framework for multi-modal human activity recognition using CCA of the three modalities: RGB, depth information and skeleton. For this purpose, we created a set of dynamic images from RGB and depth videos separately. Also, skeleton visual images were inferred from skeleton joint information. Figure. 5.1 illustrates the general pipeline of our proposed multi-modal approach.

Dynamic images are extracted from the video sequence in a way to capture spatial and temporal information among all frames. Especially, a dynamic image allows us to encode the video sequence robustly and describe the ongoing action in the video. For this purpose, we use an approximate rank pooling method as suggested in (33) to construct dynamic images. Moreover, skeleton images are constructed using 3D locations of the skeleton joints. Once the three sets of images were created, we use transfer learning from a pre-traind model to extract features from these images. Afterwards, we perform a fusion of these features using a feature-level fusion strategy based on CCA. Finally, we train a bi-directional LSTM network to recognize and classify activities in the input video sequences. In summary, our methodology is composed of four steps that we explain in detail in the following subsections.

5.3.1 Dynamic image construction for RGB and depth images

Dynamic image (DI) consists of a single image representation of a video sequence, capturing the temporal evolution of ongoing action. DIs can provide simple, powerful and efficient representations that can be used for action recognition. The concept of "Dynamic images" has been presented in (33, 34, 312). For that, the authors of (33) suggest to use an approximated rank pooling method to construct DIs. They observed that: (1) DIs focus on the motion instead of background pixels which are averaged away, (2) DIs behave differently for actions of different speeds and (3) DIs are reminiscent of some other imaging effects such as blur and panning. Similarly to (33), we use DIs to encode each video into one single image. The latter can provide us useful information on the ongoing action in the scene. We use the proposed approximated rank pooling method to calculate DIs for both RGB and depth video sequences separately. From the above and in the same spirit as (34), the video sequence is presented as a ranking function of its frames as follows:

We refer to the feature vector extracted from frame I_t by $\psi(I_t)$. So $V_t = \frac{1}{t}\sum_{\tau=1}^{t}\psi(I_\tau)$ is the average of the features extracted from frames $\{I_1, I_2 \ldots, I_t\}$ up to time t. The ranking function assigns a score $S(t|d) =< d, V_t >$ to each time increment t, where $d \in \mathbb{R}$ is a vector of parameters.

To reflect the rank of the frames in the video, d is learned as a convex optimization problem using the RankSVM formulation since later times are associated with larger scores, i.e $\forall\{q,t\}$ s.t $q > t \Rightarrow S(q|d) > S(t|d)$.

d^* is the optimizing function to the objective function given in "(5.2)" and T is the number of frames. We can see from "(5.1)" that $\rho(I_1, ...I_T; \psi)$ maps T video frames to a single vector d*. This operation of construction of d^* from T frames is called Rank Pooling.

$$d^* = \rho(I_1, ...I_T; \psi) = \underset{d}{argmin} E(d) \tag{5.1}$$

$$E(d) = \frac{\lambda}{2}||d||^2 + \frac{2}{T(T-1)} \sum_{q>t} max\{0, 1 - S(q|d) + S(t|d)\} \tag{5.2}$$

This objective function is composed of two terms: the first one corresponds to the usual quadratic regularizer of SVM while the second term serves to count how many pairs q > t are incorrectly ranked by the scoring function. In other words, it counts the number of pairs for which their associated scores are not separated by at least a unit margin.

The vector d^* contains enough information to rank all frames of the video. Similarly to (33), we apply rank pooling directly to RGB frame and depth image pixels. For that, $\psi(I_t)$ performs a component-wise non-linearity such as the square root function. As observed, d^* has the same number of elements as video frames and can therefore be used to represent the video.

Solving "(5.2)" may be computational expensive. For this purpose, we use approximated rank pooling which gives good results in practice to smooth the computation and make it faster. More specifically, the idea behind the approximated rank pooling is to consider the first step in a gradient-based optimization of "(5.2)". We then start with $d = \overrightarrow{0}$ and get a first approximated solution by gradient descent: $d^* = \overrightarrow{0} - \eta \bigtriangledown E(d)|_{d=\overrightarrow{0}} \propto - \bigtriangledown E(d)|_{d=\overrightarrow{0}}$ for $\eta > 0$ where :

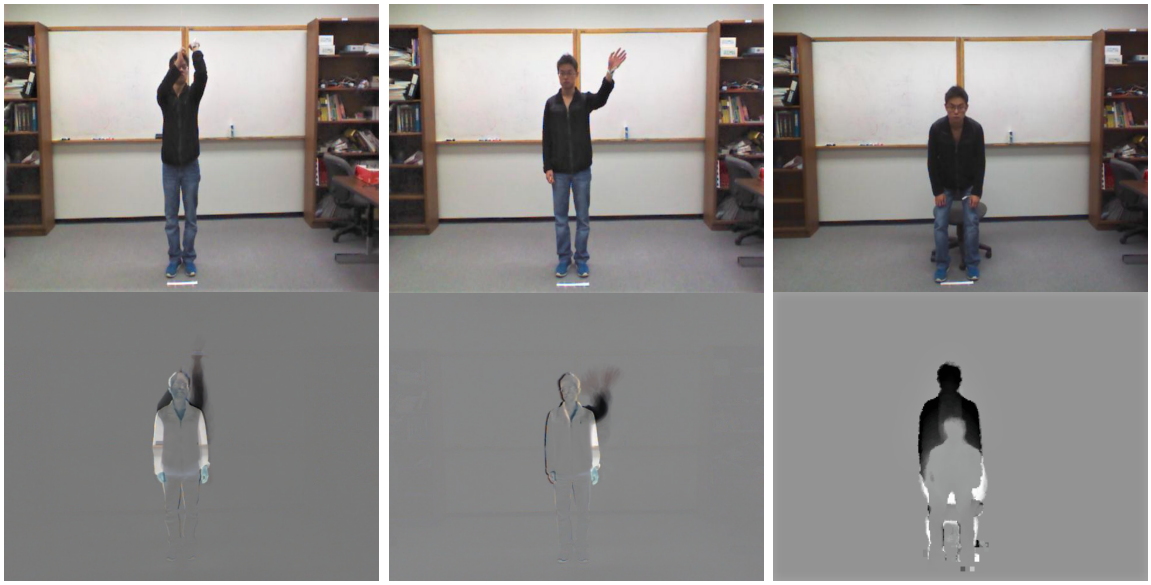
$$\bigtriangledown E(\overrightarrow{0}) \propto \sum_{q>t} \bigtriangledown max\{0, 1 - S(q|d) + S(t|d)\}|_{d=\overrightarrow{0}}$$

$$= \sum_{q>t} \bigtriangledown < d, V_t - V_q >= \sum_{q>t} < V_t - V_q > \tag{5.3}$$

So, we can extend d^* as follows, where β_t are scalar coefficients.

$$d^* \propto \sum_{q>t} < V_q - V_t >= \sum_{t=1}^{T} \beta_t V_t \tag{5.4}$$

By expanding the sum $\sum_{q>t} V_q - V_t$, each Vt appears (t-1) times with positive sign and (T-t) times with negative sign. Hence, we can deduce that $\beta_t = (t-1) - (T-t) = 2t - T - 1$.



(a) (b) (c)

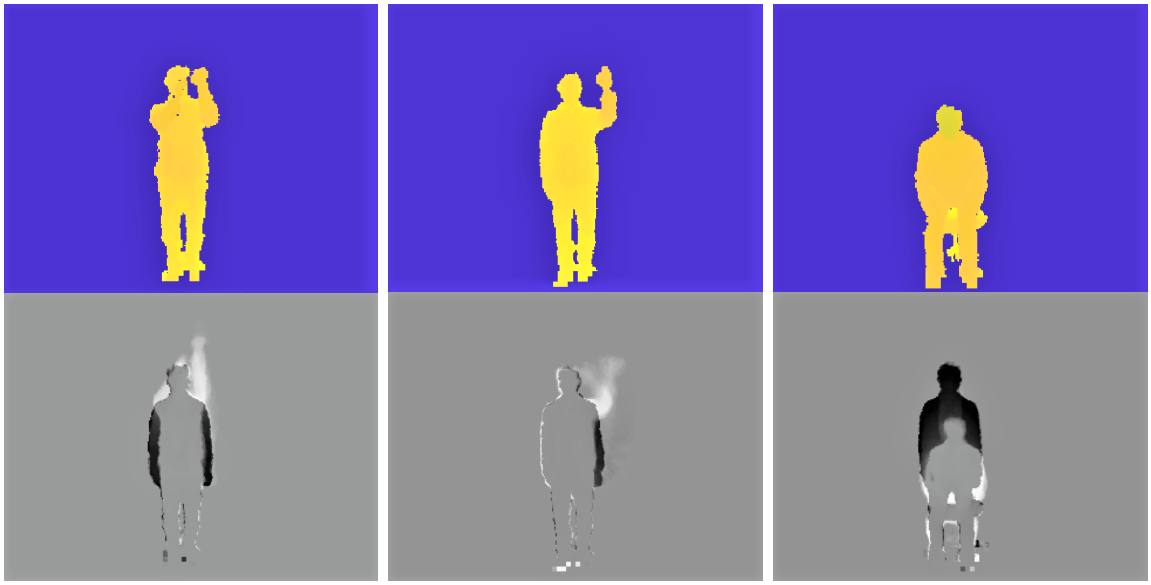
Figure 5.2: Samples of RGB video frames from the UTD-MHAD dataset (313) in the first row and their corresponding dynamic RGB images in the second row. Column (a) corresponds to a basketball shoot while the subject is waving and sitting in columns (b) and (c) respectively.

Since we already have $V_t = \frac{1}{t}\sum_{T=1}^{t}\psi(I_t)$, d^* can be written as a linear combination of the feature vector $\psi(I_t)$: $d^* \propto \sum_{t=1}^{T}\beta_t V_t = \sum_{t=1}^{T}\alpha_t\psi(I_t)$.

The approximated rank pooling is given such that the operator d^* is reduced to respect "(5.5)". So, the calculation of DIs consists in accumulating the video frames after being multiplied by α_t while $\alpha_t = 2(T-t+1) - (T+1)(H_T - H_{t-1})$ and $H_t = \sum_{i=1}^{t}\frac{1}{t}$ is the t-th Harmonic number and $H_0 = 0$.

$$\hat{\rho}(I_1, ...I_T; \psi) = \sum_{t=1}^{T}\alpha_t\psi(I_t) \tag{5.5}$$

The vectors d^*_{RGB} and d^*_{Depth} obtained from rank pooling the RGB and depth videos respectively, comprise our DIs which we call DI_{rgb} and DI_{depth}. "Figure. 5.2" and "Figure. 5.3" illustrate some examples of RGB, depth images (from UTD-MHAD dataset) and their corresponding dynamic RGB and dynamic Depth images, respectively. Columns of both images (in their order of appearance) correspond to basketball shoot, wave and stand to sit activities. We can see from these figures, that dynamic images were able to accurately summarize the execution of each of the activities as still images.



(a) (b) (c)

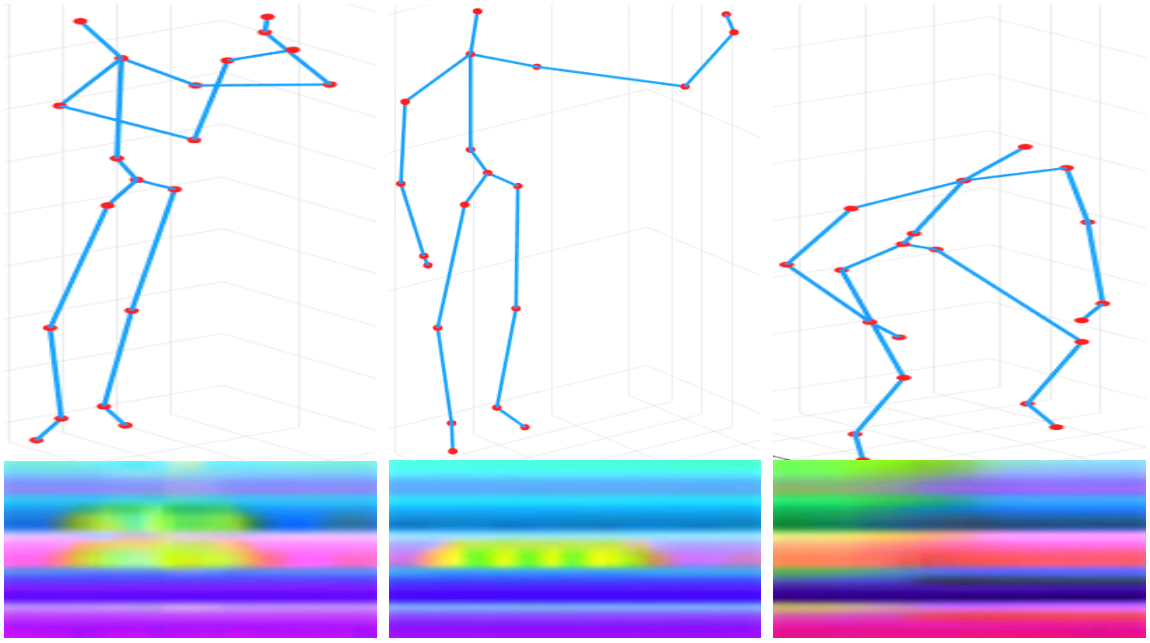
Figure 5.3: Samples of Depth video frames from the UTD-MHAD dataset (313) in the first row and their corresponding dynamic Depth images in the second row. Column (a) corresponds to a basketball shoot. In column (b) the subject is waving, and he is sitting in column (c).

5.3.2 Skeleton images from skeleton joints

Human activity recognition from skeleton information have been facing many challenges among which is: how to effectively represent spatio-temporal skeleton sequences? Moreover, retrieving features from RGB images using pre-trained models is giving very promising results in many tasks as well as for human activities recognition. Therefore, to take advantage of these models and HAR from skeleton data, we create images from skeleton sequences, then we extract discriminative features from these images using a pre-trained model. For that, and for each video sequence, we normalize the coordinates of the skeleton joints *(x,y,z)* and use them to create an RGB image which we call I_{skel}. Skeleton image allows us to track changes of each skeleton joint over time and, hence, describe the corresponding activity. "Figure. 5.4" illustrates some examples of skeleton representations from the UTD-MHAD dataset and their corresponding skeleton images. Columns correspond to basketball shoot, wave and stand to sit activities respectively.

5.3.3 Features Extraction using pre-trained models

Due to the large amounts of data needed for training an LSTM network, we extract features from our image sets {DIs_{rgb} and DIs_{depth}} using the Resnet50 model pretrained on the large Imagenet dataset. Used widely as a backbone for many computer vision tasks, it has been



(a) (b) (c)

Figure 5.4: Examples of skeleton representation from the UTD-MHAD dataset (313) in the first row and their corresponding skeleton visual images in the second row. Columns (a), (b) and (c) correspond to a basketball shoot, wave, stand to sit activities respectively.

integrated in many HAR approaches as well. It allows us to explore multiple levels of deep features by dint of its stack of layers that is composed of more than 150 layers. In addition, we use Alexnet to extract features from Is_{skel} images set. This feature extraction step is important because it provides us a strong initialisation to our feature fusion strategy compared to a straightforward use of these images. Feature vectors calculated in this step are then fused using CCA which allows us to select meaningful features.

5.3.4 Feature Fusion and activity classification

To obtain more discriminative feature vectors from our created representations, we apply a feature fusion on the extracted features from our three sets of images: the dynamic RGB images (DIs_{rgb}), the dynamic depth images (DIs_{depth}) and the skeleton images (Is_{skel}). Our feature fusion method consists of combining feature vectors of the three modalities into one single feature vector. The resulting feature vector is supposed to be more meaningful than each single aforementioned modality related feature vector. For that, and similarly to (305), we use CCA, which has been widely used for feature fusion.

Let our three feature vectors be $V_x \in \mathbb{R}^{\text{pxn}}$, $V_y \in \mathbb{R}^{\text{qxn}}$ and $V_z \in \mathbb{R}^{\text{rxn}}$ extracted from dynamic RGB images, dynamic depth images and skeleton images respectively. Each of these

vectors contain n samples. To get a representative feature vector that fuses the three vectors, we apply CCA twice. We apply CCA firstly on two vectors, for example V_x and V_y, we obtain F_1. Then, we apply CCA again on F_1 and V_z.

For each two vectors X and Y, we calculate the within-sets covariance matrices and the between-set covariance matrix that we call: $S_{xx} \in \mathbb{R}^{\mathrm{pxp}}$, $S_{yy} \in \mathbb{R}^{\mathrm{qxq}}$ and $S_{xy} \in \mathbb{R}^{\mathrm{pxq}}$ respectively and S_{yx} is the transpose of S_{xy}: S_{xy}^T. Next, we create the covariance matrix $S \in \mathbb{R}^{(\mathrm{p+q})\mathrm{x}(\mathrm{p+q})}$ as illustrated below.

$$S = \begin{pmatrix} S_{xx} & S_{xy} \\ S_{yx} & S_{yy} \end{pmatrix} = \begin{pmatrix} cov(X) & cov(X,Y) \\ cov(Y,X) & cov(Y) \end{pmatrix} \tag{5.6}$$

It is observed that understanding the correlations between X and Y using the covariance matrix S is difficult. Therefore, CCA is used to maximize the pairwise correlations given in "(5.7)" across the two data sets using Lagrange multipliers. Solution to the objective function is given by the optimizer linear combinations X^* and Y^*.

Canonical variates X^* and Y^* are defined as $X^* = W_x^T X$, $Y^* = W_y^T Y$ and $var(X^*) = var(Y^*) = 1$ where $cov(X^*, Y^*)$, $var(X^*)$ and $var(Y^*)$ are calculated using the set of equations "(5.8)".

$$corr(X^*, Y^*) = \frac{cov(X^*, Y^*)}{var(X^*)var(Y^*)} \tag{5.7}$$

$$\left\{ \begin{matrix} cov(X^*, Y^*) = W_x^T S_{xy} W_y \\ var(X^*) = W_x^T S_{xx} W_x \\ var(Y^*) = W_y^T S_{yy} W_y \end{matrix} \right\} \tag{5.8}$$

To obtain the transformation matrices W_x and W_y, one should solve the eigenvalue "(5.9)". $\widehat{W}_x$ and $\widehat{W}_y$ are the eigenvectors and Λ^2 is the diagonal matrix of eigenvalues or squares of the canonical correlations. For each equation, the number of non-zero eigenvalues (i.e. $\lambda_1 >= \lambda_2 ... >= \lambda_d$) that are sorted in decreasing order is $d = rank(S_{xy} <= min(n, p, q))$. W_x and W_y consist of sorted eigenvectors corresponding to the non-zero eigenvalues.

$$\left\{ \begin{matrix} S_{xx}^{-1} S_{xy} S_{yy}^{-1} S_{yx} \widehat{W}_x = \Lambda^2 \widehat{W}_x, \\ S_{yy}^{-1} S_{yx} S_{xx}^{-1} S_{xy} \widehat{W}_y = \Lambda^2 \widehat{W}_y \end{matrix} \right\} \tag{5.9}$$

Hence, the covariance matrix S defined above will be of the following form. Let I_d be the identity matrix and $diag(\lambda_1, ..., \lambda_d)$ be the diagonal matrix of the associated eigenvalues.

$$S = \begin{pmatrix} I_d & diag(\lambda_1, ..., \lambda_d) \\ diag(\lambda_1, ..., \lambda_d) & I_d \end{pmatrix} \tag{5.10}$$

We can observe that X^* and Y^* have non-zero correlation only on their corresponding indices and are therefore uncorrelated within each data set. Finally, we perform feature-level

fusion by concatenating the transformed feature vectors. The resulting feature vector F_1 is used to perform another time the feature-level fusion by concatenating the transformed feature vectors F_1^* and Z^* (Z corresponds to the third modality feature vector).

$$F_1 = \begin{pmatrix} X^* \\ Y^* \end{pmatrix} = \begin{pmatrix} W_x^T X \\ W_y^T Y \end{pmatrix} = \begin{pmatrix} W_x & 0 \\ 0 & W_y \end{pmatrix}^T \begin{pmatrix} X \\ Y \end{pmatrix} \tag{5.11}$$

Once our fused feature vectors are calculated, we perform activity recognition using a bi-directional LSTM network. This allows us to comprehend temporal dynamics encoded by the feature extractor (Resnet50 model for RGB and depth images and Alexnet for skeleton images) into feature maps. These feature maps are fused using CCA and fed to the classifier.

5.4 Experimental results

We evaluate our approach on the publicly available datasets UTD-MHAD (313) and NTU RGB+D (77). In the subsequent sections, we present a brief description of these datasets followed by our experimental results. We compare the recognition performance of each individual sensor modality to the performance of combining each pair of modalities and finally to the performance of fusing the three modalities. We display our performance results in Table 5.3 and Table 5.4 for UTD-MHAD and NTU RGB+D datasets.

In our LSTM model, we incorporate a bi-directional long short term memory layer. We use 600 hidden units and a feature sequence of 1000 to 1092 length. A mini-batch size of 32 samples is employed to train images of the subset and we calculate the accuracy. For the UTD-MHAD dataset, we create the training and the testing subsets using the same protocol as (313). Data from the subject numbers 1, 3, 5, 7 were used for training, while data for the subject numbers 2, 4, 6, 8 were used for testing. We report the classification accuracy on the NTU RGB+D dataset by following the action classification evaluation protocol presented in (77): cross-view evaluation, where videos from cameras 2 and 3 are used for training while videos from camera 1 are used for testing.

5.4.1 Datasets

UTD-MHAD is a multi-modal dataset (313), composed of four data modalities: RGB videos, depth videos, skeleton joint positions and inertial sensor signals. The dataset includes 861 video sequences and was recorded using a Microsoft Kinect sensor and a wearable inertial sensor in an indoor environment. It consists of 27 different actions performed by 8 subjects. Each of them repeats the same action 4 times. The Kinect sensor captures RGB sequences with a resolution of 640x480 pixels and 16bit depth sequences with a resolution of 320x240 pixels. The frame rate is approximately 30 frames per second and a time stamp for each

sample was recorded, for data synchronization. Table 5.1 summarizes detailed information of this dataset.

Table 5.1: UTD-MHAD Dataset information

UTD-MHAD Dataset	
Data modalities	RGB, Depth videos, Skeleton data and Inertial signals
Number of videos	861
Number of actions	27 different actions
Number of actors	8 subjects
Frame rate	30 fps
Acquisition device	Microsoft Kinect sensor and wearable inertial sensor

NTU RGB+D is a large-scale dataset for multi-modal human action recognition. It includes 56880 videos of 60 action classes of 40 subjects recorded in highly variant camera settings, where each action is performed twice. Three Microsoft Kinect v2 sensors were used to collect four data modalities. RGB videos are recorded in a 1920x1080 resolution and 3D skeletal information corresponds to 25 body joints. Moreover, infrared sequences and depth data are also collected and stored frame by frame in 512x424. Table 5.2 summarizes detailed information of this dataset.

Table 5.2: NTU RGB+D Dataset information

NTU RGB+D Dataset	
Data modalities	RGB, Depth videos, Skeleton data and Inertial signals
Number of videos	56880
Number of actions	60 different actions
Number of actors	40 subjects
Acquisition device	Microsoft Kinect v2 sensors

5.4.2 Results and analysis

First, we calculate the performance accuracy of each single modality. In other words, we compare activity classification from straightforward images towards newly created images (dynamic RGB, depth images and skeleton images). For both situations, we calculate the accuracy of applying LSTM on the extracted features using a pretrained model. Resnet50 and Alexnet are used as feature extractors.

As can be seen from Table 5.3 which illustrates the results of uni-modal activity recognition for UTD-MHAD and NTU RGB+D datasets, the accuracy was improved when using

our created images. For the UTD-MHAD dataset, skeleton features perform the best accuracy value for both configurations with 74.52% for skeleton joints sequences and 87.43% for skeleton images using Alexnet as feature extractor. Similarly, for the NTU RGB+D dataset, the best accuracy of 49.91% was obtained for skeleton joints sequences while dynamic depth images outperform the dynamic RGB and skeleton images with a value of 51.66%.

Table 5.3: Accuracy (%) of activity classification with LSTM of uni-modal features and features extracted (using pre-trained models) from our newly created image representations on the UTD-MHAD and NTU RGB+D datasets.

Uni-modal feature	**UTD-MHAD**	**NTU RGB+D**
RGB	51.35	39.85
Depth	37.45	45.90
Skeletal data	74.52	49.91
Our proposed images		
Dynamic RGB	72.28	41.53
Dynamic Depth	71.91	51.66
Skeleton images	87.43	50.81

Table 5.4: Accuracy (%) of activity classification using fusion of multi-modal features extracted (using pre-trained models) from our newly created image representations on the UTD-MHAD dataset and NTU RGB+D dataset respectively (DI refers to dynamic images).

Pairwise Fusion	**UTD-MHAD**	**NTU RGB+D**
DI RGB + DI Depth	85.39	60.42
DI RGB + Skeleton images	93.26	68.62
DI Depth + Skeleton images	**97.95**	**70.85**
By three Fusion		
(DI RGB + DI Depth) + Skeleton images	**98.88**	**75.50**
(DI RGB + Skeleton images) + DI Depth	92.13	73.72
(DI Depth + Skeleton images) + DI RGB	93.26	72.64

Furthermore, we calculate the recognition accuracy for each pairwise fusion and for the three features fusion. We compare in Table 5.4, the results obtained using a feature-level fusion: the Canonical Correlation Analysis on each set of features. We can see from these tables that, by combining the features from each two sets of images, the recognition accuracy was improved over that using a single modality alone for both datasets. The best results were obtained by fusing dynamic depth and skeleton images as they present complementary temporal features. We achieve for that an accuracy of 97.95% for the UTD-MHAD dataset and 70.85% for the NTU RGB+D dataset.

Table 5.5: Comparison of the proposed method with previous methods on UTD-MHAD Dataset.

Method	Accuracy %
Decision Fusion Using LOGP (314)	88.40
Depth + inertial data fusion + CRC classifier (313)	79.10
5-CNN fusion of skeleton images (315)	95.38
fusion with CCA and KELM (316)	97.91
DI RGB + DI Depth + Skeleton images + LSTM (Ours)	**98.88**

Table 5.6: Comparison of the proposed method with previous methods on NTU RGB+D Dataset.

Method	Accuracy %
Deep RNN (77)	64.09%
Deep LSTM (77)	67.29%
Joint trajectory maps + CNN (317)	75.20%
Part-aware LSTM (77)	70.20%
DI RGB + DI Depth + Skeleton images + LSTM (Ours)	**75.50%**

Fusing the three modalities has as well improved the recognition accuracy over that using single modalities or pairwise modalities. The order of fusing features was also investigated and the results demonstrate that when changing this order, the accuracy is also improved. We obtain an accuracy of 98.88% for fusing RGB and depth dynamic images and then fusing the resulting vector with skeleton images feature vector for the UTD-MHAD dataset and an accuracy of 75.50% for the NTU RGB+D dataset for the same configuration.

Table 5.5 presents a comparison of the results of our method to the state-of-the-art on the publicly available UTD-MHAD dataset. We can see that our method outperforms all the previous methods of feature fusion on the UDT-MHAD dataset. Again, Table 5.6 illustrates a comparison of our results with the state-of-the-art results on the NTU RGB+D dataset. Our results can achieve high recognition accuracy and compete with the state-of-the-art HAR approaches such as (317) and (77). However, we still can improve the results by enhancing the CCA fusion strategy.

5.5 Conclusion

We presented here a vision-based multi-modality fusion approach for HAR. RGB, depth images and skeleton joint data are used to construct RGB dynamic images, depth dynamic images and skeleton images, respectively. These constructed visual images are then employed to generate features using pre-trained models that allow us to retrieve meaningful features

from the image sets. Afterward, for each video sequence, a feature fusion strategy based on the Canonical Correlation Analysis is carried out to select highly discriminative features from our three feature vectors. The resulting feature fusion vectors are then fed to a bi-LSTM network in order to recognize and classify activities. We evaluate our approach on the publicly available UTD-MHAD and NTU RGB+D datasets and record recognition accuracy for each single modality, fusion of each pair of modalities and fusion of three modalities. Our experiments show that the results of our proposed approach can achieve high recognition accuracy and outperform the state-of-the-art results for both datasets. In the future, we can explore other fusion schemes and integrate some data augmentation methods to improve the performance of our proposal. Besides, we believe there is also a room for further improvement on the recognition accuracy achieved by NTU RGB+D dataset throughout a more fine-gained optimization of the parameters of the underlined LSTM model.

6

Vision-based Fall detection using body geometry and pose estimation

6.1 Introduction

Performing regular daily life activities by the elderly population can be affected by many serious health issues among which fall and its resulting injuries are the most frequent. This is mostly experienced by the elderly because of the natural phenomenon of brain cells death, which impact the functioning of the nervous system and, thereby, the cognitive capability of the individual (260). This results in problems with movement and safety issues. Moreover, falling is due to other inherent factors such as age-related biological changes, neurological disorders, physiological health profile and environmental conditions (260). Authors in (260) presented a detailed study of the different factors that may lead to falls in elderly population. In general, falls can result from sudden loss of balance, stability, dizziness, or vertigo during daily life movements. It can also be caused by chronic diseases, cognitive impairment, using a walking aid or multiple medications, gait and visual deficit (261). Falling is an abnormal human activity that occurs infrequently and unpredictably. It is defined by (262) as an event resulting in a person coming to rest inadvertently on the ground or the floor or any other lower level. Falls can occur in many different ways such as backward fall, due, for example, to a slippery floor, forward fall caused by tripping, side-way fall due to miss-stepping and straight-down fall due to fainting (268).

Intelligent video surveillance (camera-based approach), through a continuous monitoring by an operator, is a simple way to detect fall and trigger appropriate actions. Nevertheless, privacy concerns and often reluctant users to wear wearable devices, raises the importance of automatic detection technology. In this regard, the development of automatic detection of fall using video sequences is challenging due to constant changing of room lighting conditions and the variety of daily activities that may resemble to a fall, which causes inherent difficulties

to image processing tasks. Therefore, the accuracy of existing video based FD systems is often moderate to low, which call for further research in this issue. This partly motivates the current contribution. Especially, new features are put forward by observing that the vector formed by the head and the center hip of the body is aligned horizontally and in parallel to the ground during a falling posture, while it is perpendicular to the ground axis in a sitting or standing posture. A novel machine learning-like approach for FD from video sequences is devised. Our approach relies on calculating the angle and the distance between the vector formed by the head and the body's center hip and the horizontal axis passing through the body's center hip. For each video sequence, we calculate the aforementioned angle and distance for all frames. The computed angles and distances form the new feature sets that characterize the video sequences. We train an LSTM and a TCN networks on these features to recognize fall and non-fall activities. Furthermore, we construct new images using these angle and distance sequences so that each video sequence is represented by one image of its corresponding angles and distances. Then, a two-class SVM is trained on these images to detect fall and non-fall activities. We use the Le2i dataset (300) and the UR FD dataset (302) to evaluate the performance of our method by implementing a cross-dataset evaluation. For this purpose, we compare the results of using UR-FD for training and Le2i dataset for testing, to its reciprocal (i.e., Le2i for training and UR-FD for testing). Different performance metrics have been employed to quantify the quality of the designed classifiers in detecting falls. The experimental results indicate the feasibility of the developed approach while achieving state-of-art performance in terms of FD accuracy.

The rest of this chapter is organized as follows. First, we briefly provide background and previous research related to vision-based fall detection (FD) in Section 6.2. Section 6.3 outlines our approach. Then, we describe and discuss in Section 6.4 the experimental results of our proposal on the publicly available datasets. Finally, we conclude and set future directions for fall detection in Section 6.5.

6.2 Related Work

It is acknowledged that fall is one of the major public health problems in the world that should be carefully addressed. It is ranked second among the leading causes of accidental or unintentional injury deaths (262), especially for elderly population. Tuner et al. (318) predict that Fall will cause 45 billion Euros yearly in Europe and around 50 billion dollars in USA according to Florance et al. (319). This testifies on the huge impact of Fall on both individual and societal scale, which raises the importance of prior detection or at least rapid intervention, whenever needed, to reduce the risk of complication. In this respect, Fall Detection (FD), Fall Classification (FC), and Fall Prediction (FP) are recognized as important research directions

in the study of fall and are among the hottest topics in health-care national policy as well. The availability of efficient methods to identify and, possibly, predict fall occurrence can have a substantial public impact since it may significantly minimize damages, enable useful medical assistance, and provide daily health care to vulnerable populations (265). Moreover, missing to identify falls can expose individuals to serious health and safety risks. This has a noticeable effect on individual autonomy, independence, and life quality. It is to note that experiencing fall or near-fall events (such as missteps or stumbles) may lead to Basophobia, also called fear of falling (265, 271). This syndrome can cause many other disorders such as lack of mobility and independence, and/or social isolation (267). Several FD systems have been put forward by both research community and commercial entities (320) using various technologies and at various level of maturity. Such technologies can be classified into three streams according to the employed sensors: ambience device, wearable device and camera-based (261).

6.3 Proposed Method

The central key in our developed methodology consists in identifying relevant features that can genuinely distinguish fall from non-fall activities in 2D representation. In this respect, we noticed that, when a person is sitting or standing, the head and the center hip form a vector which is perpendicular to the horizontal axis passing through the center hip, as illustrated in "Figure. 6.1 a" and "Figure. 6.1 b". The horizontal axis is defined as a straight line parallel to the *X_axis* and passing through the center hip. In contrast, when a person is in a lying or a falling posture, this vector is approximately aligned and parallel to the horizontal axis of the body's center hip. Besides, sitting slumped to one side leads to forming an angle of around 120° or 45° between the mentioned vector and the horizontal axis, as shown in "Figure. 6.1 c" and "Figure. 6.1 d". The angle value depends on the degree of slump sitting. However, the posture is considered lying or falling when this value is close to 0° or 180°, as shown in "Figure. 6.1 e".

To illustrate our approach mathematically, we refer to the head centroid by the point $H(x_h,y_h)$ and to the center hip of the body by the point $B(x_b,y_b)$. Let $\vec{U}$ be the vector from B to H. Similarly, let $\vec{V}$ be the vector joining the point B to the point $C(x_c,y_c)$. The point C is defined such that $x_c > x_b$ and $y_c = y_b$. The red cross in Fig.6.1 refers to the head (point H) and the blue cross refers to the center hip of the body (point B).

Relying on the above observation, we calculate, for each video frame, the angle α formed between $\vec{U}$ and $\vec{V}$, and the distance γ between the head and the center hip of the body (i.e., the magnitude of the vector $\vec{U}$). These notations are used throughout the paper. Fig.6.2 illustrates the two vectors $\vec{U}$ and $\vec{V}$, the angle α and the magnitude γ. Each video is therefore characterized by a feature vector containing the sequence of the computed angles (α) and

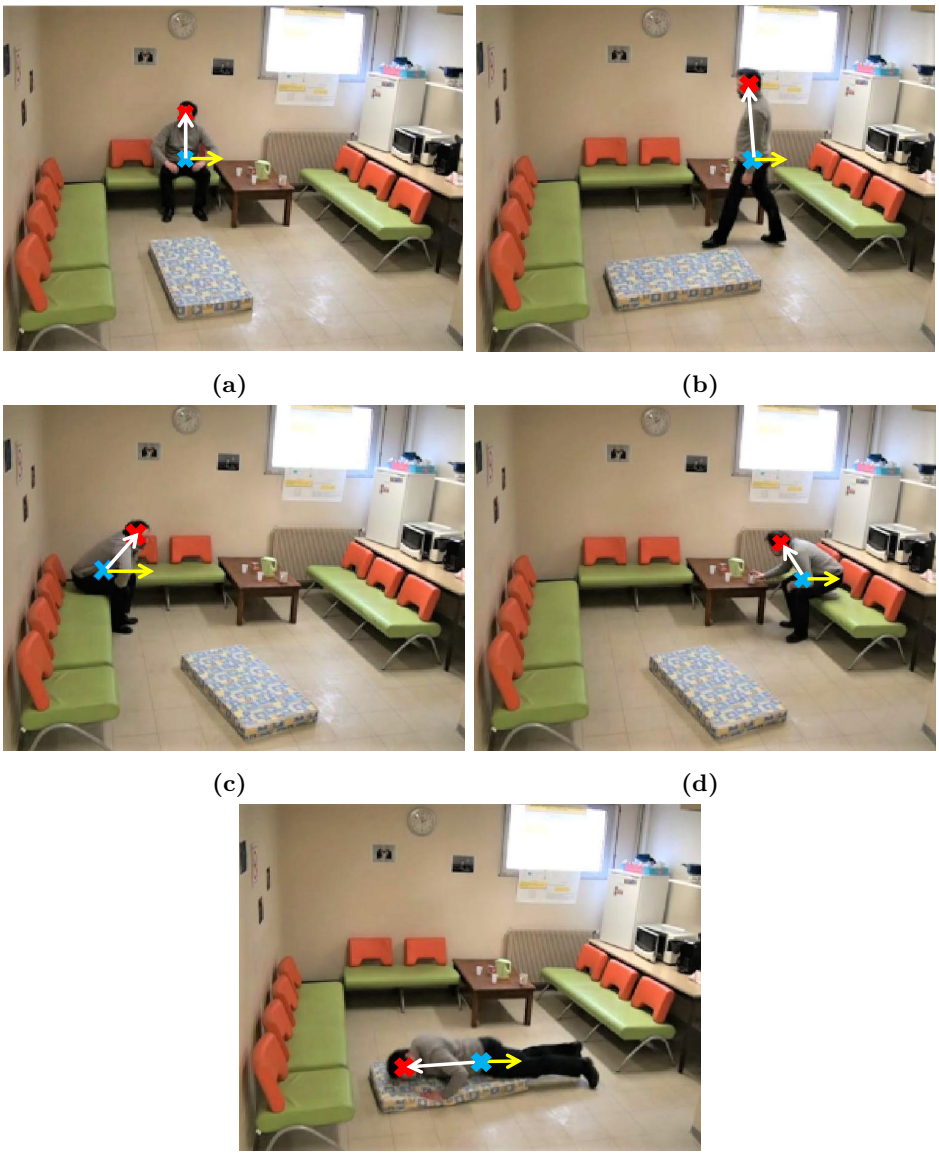


Figure 6.1: Samples from the Le2i fall detection dataset (300) representing the angle α in (a) sitting, (b) standing, (c) bending to the right posture and (d) bending to the left and finally, (e) falling postures. The value of α is around 90° in (a), (b), around 120°in (c), around 45°in (d), and around 180° in (e). α is calculated between the white and the yellow vectors.

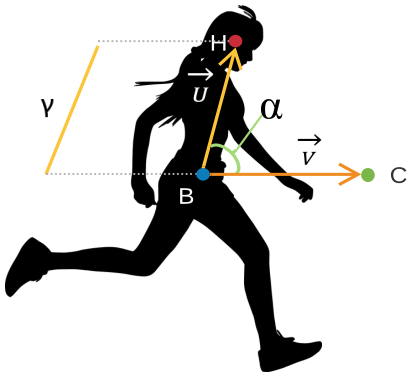


Figure 6.2: Mathematical representation of our method

distances (γ) of all its frames. The first step of our approach consists of down-sampling the videos to reduce the number of frames by keeping only those that contain motion. This helps us to reduce both the computational and the man-power burden effort in the next steps when the annotation is conducted manually. This step is omitted when the annotation is performed automatically where the whole video sequence is processed. Then, we use these frames to annotate the head and the body parts, distinguishing the manual annotation process (of the head and of the body part of each frame), and the automatic annotation process through a pose estimation algorithm that uses a deep learning approach whose detail is provided later on. We propose to use either the manual or automatic annotation to localize both points (the head and the center hip of the body) and the choice between two alternatives can be either manual or system requirement. For instance, if the system should be made real-time, the automatic annotation is more convenient. However, if the most important features of the FD system are the precision and the reliability, it is better to opt for the manual annotation. We then calculate the centroid and the center hip of the head and body part, respectively. Next, we calculate the angle α and the distance γ. Next, we represent each video with a sequence of angles α_i and a sequence of distances γ_i, where i represents the frame's index. Once these sequences are created for each video, we distinguish two scenarios. In the first scenario, we train an LSTM network on these features and classify the video sequences accordingly into fall and non-fall cases. Here 'non-falls' would stand for all other daily living activities that are not falls. We also devised a data augmentation strategy that handles the potential mismatch of input size to the LSTM network. For this purpose, we performed a simple padding task where the feature vector (angle and distance value) of the first frame is duplicated and concatenated to yield the same dimension as the largest sequence. In the second scenario, we create a set of images that is fed later to some pre-trained network to extract distinctive features. The extracted features are then trained with a two-class SVM

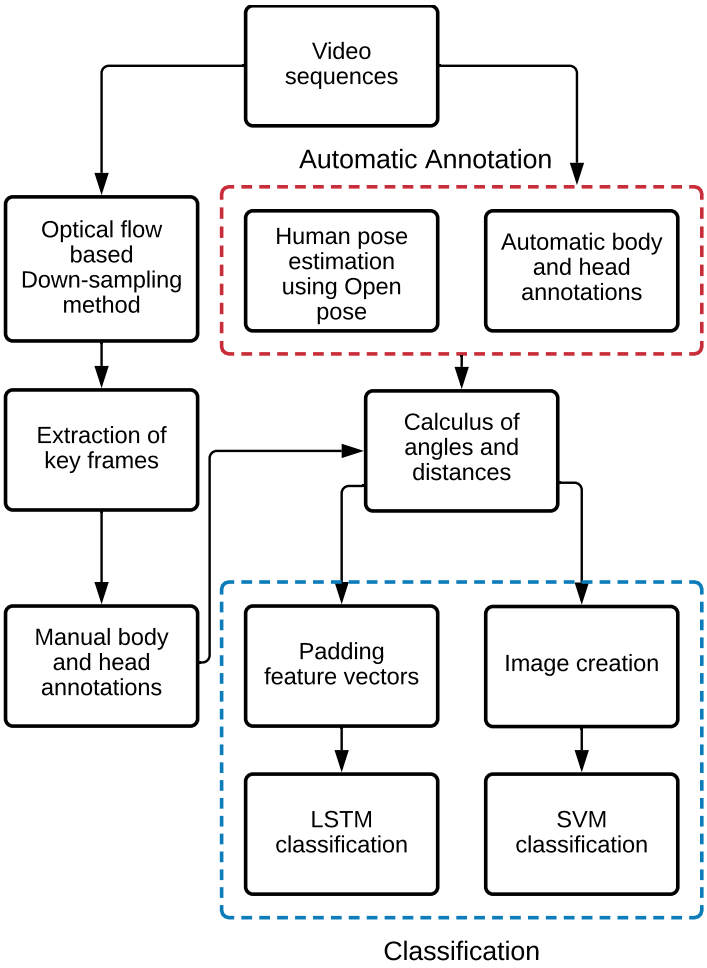


Figure 6.3: The pipeline of our proposed fall detection approach

classifier to detect falls from daily life activities. Finally, the results of the classification with SVM, LSTM and TCN are compared. The overall approach is summarized in four steps as follows (see Algorithm. 1 and Fig. 6.3).

6.3.1 Step 1: Down-sampling the videos

The length of the Le2i dataset video sequences vary from 30 seconds to 4 minutes, with a frame rate of 25 frames per second, yielding video sequences of more than 1000 frames. This renders any manual annotation task very exhausting. Therefore, for the purpose of reducing the intensive labor work, instead of using all the frames, we down-sample each video (note that this only applies to manual annotation task not in the case where automatic annotation were used). Inspired by (321) where a motion analysis based video summarization technique using an optical flow has been presented, we use optical flow (OF) to down-sample our videos. For this purpose, we exploit the Horn-Schunck method to estimate the movement in video

Algorithm 1: Calculate angles and distances

(I) INPUT: video_sequence, method, bool;

(II) **if** *method = 'Manual'* **then**

- Use optical flow to extract significant frames;
- Use key frames to down-sample the video;
- Manually annotate head and calculate the centroid H;
- Manually annotate body and calculate the center hip B;

else if *method = 'Automatic'* **then**

- 1 - Extract feature maps;
- 2 - Predict 2D confidence maps of body part locations;
- 3 - Predict 2D vector fields of part affinities that describe the degree of association between body parts;
- 4 - Produce 2D key-points for humans in the scene;
- 5 - Get locations of the point head;
- 6 - Use the locations of right knee, and left knee points to calculate the center hip of the body;

else

- print('Available methods are Manual or Automatic');

end

(III) **for** *frame in video_sequence* **do**

- Create point C with Xc > Xb (for example Xb+50) and Yc = Yb;
- Create vectors $\vec{U}$ and $\vec{V}$;
- Calculate angle between vectors $\vec{U}$ and $\vec{V}$ using the cosine law ;
- Calculate distance between vectors $\vec{U}$ and $\vec{V}$;
- Save frame, angle, distance;

end

sequences. Especially, we observed that falls, in general, occur fast and, therefore, can be characterized by a significant motion change among frames. Accordingly, we keep track only of those frames that bear important motion change and construct a re-sampled video sequence using these frames, with a rate of 25 frames per second. It is to note that the new re-sampled videos contain fewer frames than the original videos. For example, a video of 1607 frames is reduced to 578 frames, while another video of 1283 frames is downsampled to 583 frames. The downsampling rate is variable and depends on the mean values of the optical flow components of each video. More specifically, to automate this downsampling process, we first estimate the optical flow among all the frames using the Horn-Schunck method that consists in resolving the constraint: $I_x.u + I_y.v + I_t = 0$. Where I_x, I_y, I_t are the spatiotemporal image brightness derivatives, while u and v correspond to the horizontal and the vertical optical flow components, respectively. Then, we calculate the mean of both

horizontal *(Vx)* and vertical *(Vy)* components of the optical flow, which we call $meanVx$ and $meanVy$, respectively. Subsequently, the mean squared normalized error performance (MSE) is computed to estimate the similarity between the horizontal\vertical components of optical flow of each frame and the mean value of horizontal and vertical components of optical flow, respectively ($similarityVx$ and $similarityVy$). "(6.1)" demonstrates how to calculate such similarity using the MSE values.

$$similarityVz_i = \frac{1}{P}.\sum_{p=1}^{P}(Vz_i(p) - meanVz)^2 \tag{6.1}$$

Where z refers to either x or y component. P refers to the pixels of the frame, while i corresponds to its index and p stands for a particular pixel of the frame i. Frames that have a similarity $similarityVx_i$ (resp. $similarityVy_i$) above or equal their mean similarity $meanSimVx$ (resp. $meanSimVy$) are preserved while others are removed. The mean similarity $meanSimVx$ (resp. $meanSimVy$) is computed as the average similarity $similarityVx_i$ (resp. $similarityVy_i$) across all the frames. This process constructs the re-sampled videos. Besides, the maintained frames should respect the conditions given by "(6.2)".

$$\begin{Bmatrix} similarityVx_i >= meanSimVx \\ similarityVy_i >= meanSimVy \end{Bmatrix} \tag{6.2}$$

This re-sampling process (see Algorithm. 2) reduces the number of frames of each video to less than half of the original number of frames of the given video, which, in turn, facilitates the subsequent task of manual annotation.

6.3.2 Step 2: Body and head annotations

Once the videos are re-sampled, and since our work uses features extracted from the body geometry, we describe below the manual and automatic annotation of individual's head position and body part in each video frame.

6.3.2.1 Manual annotation

The annotation of each frame contains the frame's index, the localization of the head part in the video frame in terms of the approximated bounding box (according to human visual perception capability), and the coordinates of the head centroid (approximated by the center of the bounding box). Similarly, to annotate the human body part in the video frame, discrete points are manually assigned over the contour of the body with the help of online annotation tool, which is then used to estimate its centroid that corresponds to the center hip. Fig. 6.4 a and Fig. 6.4 b illustrate the manual annotation of the body and the head part. The blue

Algorithm 2: Down-sampling videos

```
(I) INPUT: video_sequence, method, bool;
framex = {}, framey = {};
for frame in video_sequence do
    Estimate optical flow;
    Calculate horizontal and vertical components Vx and Vy ;
    VxFrames(frame,:,:) = Vx;
    VyFrames(frame,:,:) = Vy;
end
Calculate mean(VxFrames) and mean(VyFrames);
for frame in video_sequence do
    similarityVx(frame) = mse (VxFrames,Vx);
    similarityVy(frame) = mse (VyFrames, Vy);
end
Calculate mean(similarityVx) and mean(similarityVy);
for frame in video_sequence do
    if similarityVx(frame) >= mean(similarityVx) & similarityVy(frame) >=
     mean(similarityVy) then
        framex.append(frame);
    frames = unique(framex,framey, 'sorted');
end
```

point corresponds to an estimation of the center hip of the body in (a) and the centroid of the head in (b). The samples were taken from the Le2i dataset. The maintained frames from the previous re-sampling process are used here to annotate the video sequences. The shapes surrounding the head and the body are drawn manually and their centroids are then calculated.

6.3.2.2 Automatic annotation

For the automatic annotation, we used a pose estimation model that relies on the pre-trained Caffe model, which won the CoCo keypoints challenge in 2016 (322). The position and orientation of the human body are tracked across frames by detecting keypoints that correspond to important parts of the body and localizing individual's major joints. More specifically, the pre-trained model was trained on the multi-person dataset MPII (323) that produces 15 points as illustrated in Fig. 6.4 c and outputs the confidence score and affinity maps. The detection of the keypoints proceeds in three stages. First, the image inputs are fed to ten first layers of a VGG network to extract feature maps. Next, a 2-branch multi-stage CNN is implemented where 2D confidence maps of body part locations and 2D vector fields of

(a) (b)

(c)

Figure 6.4: Samples from the Le2i FD dataset representing: First row - the manual annotation of a) the center hip of the body and b) the head; Second row - the points produced using the pre-trained model trained on the multi-person dataset MPII (automatic annotation).

part affinities that describe the degree of association between different parts are predicted in the first and second branch, respectively. Finally, the 2D keypoints for human bodies in the scene are produced by parsing both confidence and affinity maps using greedy inference. Once we get the locations of the key points at each frame of the video, we focus only on 3 points: head, right knee, and left knee. The point head is used to localize the head while the two other points are used to calculate the center hip of the body. Note that the pose estimation is used to determine the localization of the head and the center hip of the body only and not for the feature extraction.

6.3.2.3 Angle and distance calculus

The head centroid and the center hip of the body are used to calculate their associated distance γ and the angle α between the vector $\vec{U}$ and the vector formed by the horizontal axis corresponding to the x coordinate of the center hip called $\vec{V}$.

For the angle calculus, we can compute its cosine value and deduce the corresponding angle. The cosine is computed using the law of cosines, and the Euclidean norm is used to calculate the magnitude of vectors. "(6.3)" highlights the process of calculating the cosine of the angle α. $(\vec{V} - \vec{U})$ refers to the vector between the head centroid and the axis point C and $\left\|\vec{X}\right\|$ is the Euclidean norm of the vector $\vec{X}$.

$$cos(\alpha) = \frac{-\left\|(\vec{V} - \vec{U})\right\|^2 + \left\|\vec{U}\right\|^2 + \left\|\vec{V}\right\|^2}{2.\left\|\vec{U}\right\|.\left\|\vec{V}\right\|} \tag{6.3}$$

We, therefore, calculate the distance between the head and the center hip of the body (magnitude of the vector $\vec{U}$) among all the video frames using the Euclidean norm for both

manual and automatic annotations as shown by 6.4.

$$\left\|\overrightarrow{U}\right\| = d(B,H) = \sqrt{(x_b - x_h)^2 + (y_b - y_h)^2} \tag{6.4}$$

6.3.3 Step 3: Feature extraction

As illustrated in Fig. 6.3, we discern two scenarios depending whether macro-images have been used to infer feature vectors or not. More specifically, in the first scenario, we construct our feature vectors by concatenating angles and distances, while in the second scenario, we create macro image features using angles and distances. Besides, the angles and the distances are calculated between the vectors $\vec{U}$ (white vector in Fig. 6.1) and $\vec{V}$ (yellow vector in Fig. 6.1) among all frames of the re-sampled videos. The details of both scenarios are provided in the subsequent subsections.

6.3.3.1 Padding feature vector

In the first scenario, each video is characterized by the feature vector illustrated by "(6.5)", where i is the index of the video frame.

$$V = \{[1, \alpha_1, \gamma_1], [2, \alpha_2, \gamma_2], [3, \alpha_3, \gamma_3] ... [i, \alpha_i, \gamma_i]\} \tag{6.5}$$

Since the video sequences do not contain the same number of frames, these feature vectors are of different lengths and could not be fed directly to the classifier which requires fixed input size. For that, we perform a padding strategy (see the first 'if padding' block from Algorithm. 3) that allows us to enforce the same vector size whose length is set to the maximum value of all vectors' lengths. Therefore, each vector is extended to the new length by adding new value to its beginning. Besides, to avoid random allocation, the new components resulting from the feature vector augmentation are assigned 'angles and distance' values corresponding to the first frame. This is due to the fact that augmenting the feature vector by duplicating the features of the first frame, which actually encodes the ongoing action in the first fame, cannot alter the actions performed by the individual in other frames. Feature vectors are then fed to an LSTM classifier for FD. Besides, features obtained from both manual and automatic annotations are considered and proceeded in the same manner. To illustrate this process, let us consider a video M characterized by:

$$M = \{[1, \alpha_1, \gamma_1], [2, \alpha_2, \gamma_2], [3, \alpha_3, \gamma_3] \dots [K, \alpha_K, \gamma_K]\}$$

Let us refer to the total number of frames by K and the maximum value of all video lengths by *Max*, where $K <= Max$. We add (*Max*-K) elements of value [α_1,γ_1] at the beginning of M, so the vector M becomes:

$$M = \{[1, \alpha_1, \gamma_1], [2, \alpha_1, \gamma_1], \ldots, [Max - K, \alpha_1, \gamma_1], \\ [1 + Max - K, \alpha_1, \gamma_1][2 + Max - K, \alpha_2, \gamma_2], \\ [3 + Max - K, \alpha_3, \gamma_3] \ldots [K, \alpha_K, \gamma_K]\}$$

Algorithm 3: Padding and classification

```
(I) INPUT: Feature_sequences, padding, maxSequence;
(II) if padding then
    for sequence in Feature_sequences do
        Features = Feature_sequences(sequence);
        firstFrame = Features(1,:);
        if length(Features) < maxSequence then
            remainingFrames = maxSequence - length(Features);
            newFeatures = zeros(remainingFrames,3);
            for i in remainingFrames do
                newFeatures(i,:) = firstFrame ;
            end
            newFeatures = [newFeatures,Features];
            Save newFeatures;
    end
    Prepare training and testing sets from newFeatures;
    Classify with LSTM;
else
    for sequence in Feature_sequences do
        Features = Feature_sequences(sequence);
        Concatenate (Features(1,:),Features(2,:),Features(3,:)) to create new image;
    end
    Prepare training and testing sets from new images;
    Transfert learning to extract features from these images;
    Classify with SVM or TCN;
end
```

6.3.3.2 Macro-Image feature

Similarly, in the second scenario, we concatenate the angles and the distances to construct the set of macro images (RGB images). The newly created RGB image for the video sequence V is thereby constructed using the following feature vector:

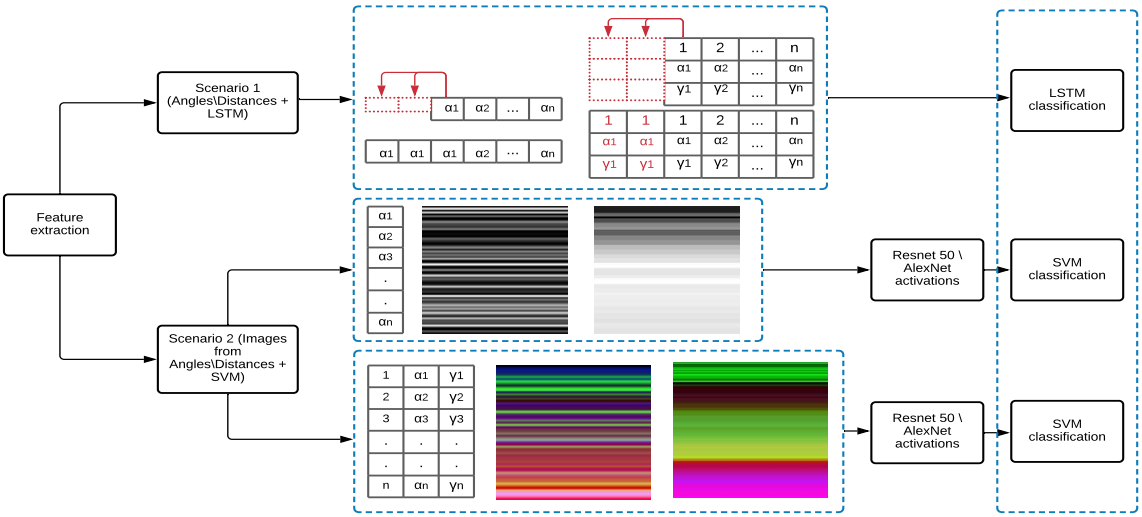


Figure 6.5: Our padding strategy followed by the feature extraction process and classification step. In the first scenario, the angles and the distances of the first frame are used to fill out the empty elements of the (augmented) feature vectors, which are then fed to an LSTM classifier. In the second scenario, the angles and the distances are used to create images which are fed firstly to a pre-trained model to extract significant features and, then used to train an SVM classifier.

$$V = \{[1, \alpha_1,\gamma_1], [2,\alpha_2,\gamma_2], [3,\alpha_3,\gamma_3] \ldots [i,\alpha_i,\gamma_i]\}$$

Where values of i build the first channel, α_i build the second channel, while γ_i make up the third channel.

Each video sequence is therefore characterized by an image from the macro images set. Accordingly, the macro image encodes the angle sequences and the distance sequences taking into account the temporal aspect of the video illustrated by the first channel (the video frames). This set of images is constructed using features extracted from 'angles and distances' using both manual and automatic annotations. Fig. 6.5 illustrates step 3 and step 4.

6.3.4 Step 4: Classification

We train a Long-short-term memory (LSTM) network using the sequence of both angles and distances in the first scenario to detect fall and non-fall cases. Besides, to make the learning faster, we construct our LSTM model using a bi-LSTM layer that allows us to access data in both forward and reverse directions. Fig. 6.6 highlights our LSTM architecture.

To detect falls in the second scenario, we extract distinctive features from our set of feature images using a pre-trained model. In our approach, we use activations of the Resnet50 and the AlexNet networks as our features. Then, we feed them to a two-class SVM classifier to distinguish between falls and daily life activities.

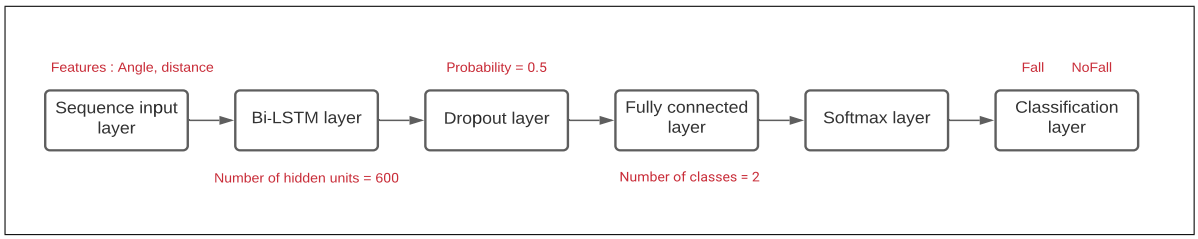


Figure 6.6: Our LSTM architecture for classifying falls using our calculated angles and distances as input sequences.

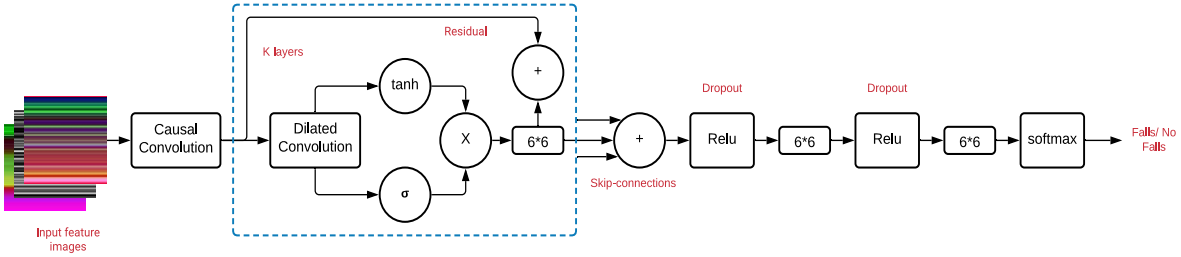


Figure 6.7: The TCN architecture (324) used for classifying falls using our calculated angles and distances as input sequences.

Furthermore, we train a TCN classifier (as illustrated by Fig. 6.7) to distinguish between falls and no falls activities from our newly created feature images. For the TCN architecture, we use 20 filters in the convolutional layers with a kernel size of 6 for each. Also, one stack of residual block is used in addition to a dropout rate of 0.05, a 'Relu' activation function and the batch normalization is exploited.

6.4 Experimental Results and Discussion

Fall is a kind of unpredictable action that occurs infrequently. Due to the rarity of the occurrence of falls, most existing FD datasets are set up by simulated fall data. The lack of such benchmark datasets and real fall data makes the evaluation process of FD systems hard and less convincing. We evaluate our approach on the publicly available Le2i FD (300) and the UR FD (302) datasets. In the subsequent section, we present these two datasets and the evaluation metrics employed to evaluate our proposal's performance and, finally, the experimental results we obtained.

6.4.1 Experimental setup

We first evaluate the results obtained from our SVM and LSTM models trained on our extracted features. Next, we evaluate the features extracted from our constructed images using the Resnet50 model to those extracted using the AlexNet model. Then, we provide results of FD using automatic annotation of the head and the center hip of the body. In

addition, we evaluated the performance of the TCN network on both; extracted features and our created feature images. Finally, we perform a cross-dataset evaluation of our FD system using the Le2i and the UR FD datasets for training and testing respectively and vice-versa.

To be able to compare our work to previous works on the same datasets and in the same way as (287), we apply a k-fold cross-validation to our LSTM and SVM models with k=5. The UR FD and the Le2i datasets were randomly split into five equal size subsets. At each iteration of the five iterations, we compose the training and testing sets with four subsets and one set. We use a data augmentation process to transform training and test sets with an optional pre-processing stage such as resizing, which helped us to resize the images of the data-store to make them compatible with the input size of the pre-trained model. Therefore, at each epoch, the training set is modified slightly to get a better result and avoid overfitting. The results are computed across the combination of all the iterations.

We calculate the Recall (or sensitivity), Precision, Specificity, Accuracy, and F1_measure for each configuration of features. We calculate the false positive rate as $FPR = 1 - specificity$ and the false negative rate as $FNR = 1 - recall$.

6.4.2 Datasets

The Le2i fall detection dataset contains 221 videos of 131 falls and 90 daily life activities (ADL). The different activities are recorded by a single fixed camera with a frame rate of 25 frames/s and a resolution of 320x240 pixels. All the activities are simulated by several actors and are gathered at four different locations: Home, Office, Coffee room and Lecture room. The dataset illustrates many difficulties of realistic video sequences of an elderly home or office such as variable illumination and occlusion. The manual annotations of 191 videos were given, with extra information representing the ground-truth of the fall position and the localization of the body in the image sequence. Table 6.1 gives detailed information of this dataset.

Table 6.1: Le2i Fall Detection Dataset information

Le2i Fall Detection Dataset	
Locations	Home, Office, Lecture room, Coffee room
Number of fall videos	192
Number of ADL videos	57
Segment length	30 s – 4 mins
Type of data	Simulated data
Frame rate	25 fps
Acquisition device	Single fixed camera

The UR fall detection dataset contains 70 (30 falls + 40 activities of daily living) sequences (302). Two Microsoft Kinect cameras were used to record fall events from two different perspectives where ADL were recorded with only one camera. This results into 60 fall sequences and 40 non-fall activities. Table 6.2 gives detailed information of this dataset.

Table 6.2: UR Fall Detection Dataset information

UR Fall Detection Dataset	
perspectives	Two perspectives for fall events and one for ADL
Number of fall videos	60
Number of ADL videos	40
Type of data	Simulated data
Frame rate	25 fps
Acquisition device	Microsoft Kinect camera

6.4.3 Experiment results

6.4.3.1 Evaluation on the Le2i FD dataset

For the Le2i dataset, we achieved a Recall score of 100% for the set of features consisting of angles and distances and trained on the LSTM network. Beside, the outcomes of LSTM training are by far better than SVM training on the same feature set. This can be justified by the high-capacity of the deep learning models to extract significant features and classify them accordingly.

Table 6.3 illustrates the results obtained for the set of images constructed from (angle + distance) with manual annotation using the activations of the Alexnet and the Resnet50 models as well as the results of training SVM and LSTM on these features directly for the Le2i dataset. Likewise, Table 6.4 summarizes the results obtained for the same set of images using automatic annotation approach. Notice that LSTM results were found to outperform those obtained using SVM classifier, and better accuracy, precision and specificity scores were obtained by applying the TCN to the same set of features. Besides, the best accuracy, precision, recall, F_score and specificity values (88.9%, 90.0%, 90.0%, 0.90 and 100.0%) were obtained for features extracted and trained on TCN from images built using (angle + distance) based features.

6.4.3.2 Evaluation on the UR FD dataset

Similarly, for the UR FD dataset, a recall score of 100% was observed (Table 6.5) when training LSTM on the set of features consisting of (angle + distance). The recall obtained for SVM classification of images extracted from our newly created images using the AlexNet

Table 6.3: Performance results for our FD approach on the Le2i dataset using an AlexNet and a Resnet50 models for feature extraction.

Features	Acc.	Precision	Recall	F_score
Angle+Distance + SVM	0.731	0.842	0.800	0.821
Angle+Distance + LSTM	0.769	0.769	**1.000**	0.870
Feature images + AlexNet + SVM	0.885	0.952	0.909	0.931
Feature images + Resnet50 + SVM	**0.962**	**1.000**	0.950	**0.974**

Table 6.4: Performance results for our FD approach on the Le2i dataset using an AlexNet and a Resnet50 models for feature extraction and pose estimation for automatic annotation.

Features	Acc.	Precision	Recall	F_score	Spe.
Angle+Distance + SVM	0.659	0.707	0.693	0.700	0.613
Angle+Distance + LSTM	0.765	0.760	0.877	0.814	0.604
Angle+Distance + TCN	0.778	0.777	0.780	0.777	0.618
Feature images + AlexNet + SVM	0.774	0.798	0.824	0.811	0.703
Feature images + Resnet50 + SVM	0.807	0.835	0.835	0.835	0.766
Feature images + TCN	**0.889**	**0.900**	**0.900**	**0.900**	**1.000**

activations as features was better than the one obtained from the Resnet50 activations for our set of features (angles+distances). However, the accuracy and the precision values were higher when using the Resnet50 activations. Table 6.6 illustrates the results of the evaluation of our FD method on the UR FD dataset using pose estimation for automatic annotations of the head and the center hip of the body. We can notice from this table almost the same trend of outcomes as in the previous tables. Training LSTM was mostly performing better than training an SVM classifier on the features directly and TCN yields better results than SVM and LSTM in terms of Precision, Accuracy and F1 scores. Furthermore, activations of the Resnet50 provided more significant features than activations of the AlexNet. Hence classifying falls and ADLs using the newly constructed images and Resnet50 extracted features, trained on the SVM model yielded the best in terms of accuracy, precision, recall and F_score (98.6%, 100%, 97,6% and 0.988 respectively). The preceding employs the (angles and distances) features extracted using pose estimation for the automatic annotation of the head and the center hip of the body. However, the TCN network trained on our newly constructed images did not perform better than the SVM classifier as for the Le2i dataset. This can be justified

Table 6.5: Performance results for our FD approach on the UR FD dataset using an AlexNet and a Resnet50 models for feature extraction.

Features	Acc.	Precision	Recall	F_score
Angle+Distance + SVM	0.850	0.818	0.900	0.857
Angle+Distance + LSTM	0.850	0.800	**1.000**	0.889
Feature images + AlexNet + SVM	0.920	0.923	0.923	0.923
Feature images + Resnet50 + SVM	**0.960**	**1.000**	0.900	**0.947**

Table 6.6: Performance results for our FD approach on the UR FD dataset using an AlexNet and a Resnet50 models for feature extraction and pose estimation for automatic annotation.

Features	Acc.	Precision	Recall	F_score	Spe.
Angle+Distance + SVM	0.863	0.930	0.833	0.879	0.906
Angle+Distance + LSTM	0.890	0.866	0.967	0.914	0.775
Angle+Distance + TCN	0.950	0.971	0.875	0.913	0.869
Feature images + AlexNet + SVM	0.971	0.976	**0.976**	0.976	0.964
Feature images + Resnet50 + SVM	**0.986**	**1.000**	**0.976**	**0.988**	**1.000**
Feature images + TCN	0.850	0.852	0.833	0.842	0.920

by the fact that the TCN model requires more data for training to get better results whereas the UR FD dataset is a very small dataset.

6.4.3.3 Evaluation on the cross dataset

On the other hand, since our features consist of angles, distances and the generated images, a cross-dataset evaluation could be performed to estimate our FD system performances on large scale dataset. Sequences of angles and distances between the head and the center hip of the body do not contain any information which may be specific to the dataset such as illumination, actors, their clothing, background,...etc. So, the data from both datasets could be fused to construct a larger dataset for FD. Therefore, we compared the results of using the Le2i dataset for training which contains more video sequences and the UR FD dataset for testing and vice-versa. Table 6.7 outlines our findings using this cross-dataset evaluation for which we report results in terms of accuracy, precision, recall, F_score and specificity. One notices that the best results were mainly obtained by training TCN on features extracted from our newly created images using both angles and distances, in the first cross-dataset and

Table 6.7: Performance results for our FD approach using the Le2i dataset for training and the UR FD dataset for testing (cross dataset 1) and its reciprocal (cross dataset 2) with pose estimation for automatic annotation.

Features	Acc.	Precision	Recall	F_score	Spe.
Cross Dataset 1					
Angle+Distance + LSTM	0.760	0.725	**0.966**	0.828	0.450
Angle+Distance + TCN	0.662	0.671	0.596	0.578	0.414
Feature images + Resnet50 + SVM	0.810	**0.833**	0.846	0.839	0.758
Feature images + TCN	**0.820**	0.821	0.900	**0.859**	**0.847**
Cross Dataset 2					
Angle+Distance + LSTM	0.683	0.763	0.669	0.712	0.703
Angle+Distance + TCN	0.670	0.735	0.596	0.566	0.714
Feature images + Resnet50 + SVM	**0.960**	**0.952**	**0.983**	**0.967**	0.925
Feature images + TCN	0.824	0.811	0.865	0.837	**0.934**

by training SVM on features extracted from the same set of images using Resnet50 activations in the second cross-dataset. Also, using the UR FD dataset for training and Le2i for testing was giving better results than the reverse operation. This indicates that testing on many samples is essential to achieve high performance results.

From the results reported in Tables 6.4 and 6.6, we can see that using the TCN classifier for FD gave us promising results as well. Also, values of FPR and FNR (see Table 6.8) using the TCN classifier are low which testify of the reliability and robustness of our chosen features. However, we observe that SVM-based classifier yields relatively better performances for the Le2i dataset and the cross-dataset 1, which can be explained by the fact that the deep learning networks require large dataset for training while existing FD datasets were relatively small in scale, imbalanced and contain only few fall data. Especially, SVM was able to depict falls from non-falls activities more efficiently than the TCN model in the UR FD dataset and the cross-dataset 2.

Similarly, we observe from Table 6.3 and Table 6.5 that using the new images gave us better results than directly feeding the features vectors to our LSTM and SVM models. The results were by far improved by creating these images and extracting significant features from them using pre-trained models. Similarly, the false negative rate FNR and the false positive FPR rate were low. Therefore, the probabilities that a false alarm will be raised or that a fall will be missed by our system were low as well. This demonstrates the reliability of our system.

Table 6.8: Performance results in terms of False negative and false positive rates for our FD approach for the Le2i, UR FD and cross datasets (Cross dataset 1 refers to using the Le2i dataset for training and the UR FD for testing while Cross dataset 2 refers to its reverse operation) with pose estimation for automatic annotation.

Features	**FPR.**	**FNR.**
Le2i Dataset		
Angles + Distances + LSTM	0.396	0.123
Angles + Distances + TCN	0.382	0.220
Images from Angles + Alexnet + SVM	0.156	0.132
Images from Angles + Distances + TCN	0.000	0.100
UR FD dataset		
Angles + Distances + LSTM	0.225	0.033
Angles + Distances + TCN	0.131	0.125
Images from (Angles + Distances) + Alexnet + SVM	0.000	0.024
Images from Angles + Distances + TCN	0.080	0.167
Cross dataset 1		
Angles + Distances + LSTM	0.297	0.331
Angles + Distances + TCN	0.586	0.404
Images from Angles + Resnet50 + SVM	0.025	0.033
Images from Angles + Distances + TCN	0.153	0.100
Cross dataset 2		
Angles + Distances + LSTM	0.550	0.034
Angles + Distances + TCN	0.286	0.404
Images from Angles + Resnet50 + SVM	0.209	0.169
Images from Angles + Distances + TCN	0.066	0.135

Table 6.9: Comparison between performance results (in %) of our FD approach with other existing approaches on the Le2i dataset and the UR FD dataset

Approaches	**Le2i FD dataset**				**UR FD dataset**			
	Acc.	**Precision**	**Recall**	**F_score**	**Acc.**	**Precision**	**Recall**	**F_score**
Combined curvlets + HMM (325)	**97.02**	-	98.00	-	96.88	-	-	-
OF + CNN (287)	97.00	-	93.60	-	95.00	-	**100**	-
Dual-channel feature integration based FD (326)	96.91	97.65	96.51	97.08	97.33	97.78	97.78	**97.78**
SVC (327)	98.00	97.00	97.20	97.10	**99.6**	95.00	97.00	96.00
ours: **Angle + Distance + Resnet50 + SVM**	96.20	**100**	95.00	**97.40**	96.00	**100**	90.00	94.70
ours: **Angle + AlexNet + SVM**	76.90	81.80	90.00	85.70	95.00	**100**	91.70	95.70
ours: **Angle + Distance + LSTM**	76.90	76.90	**100**	87.00	85.00	80.00	**100**	88.90

The corresponding values of FPR and FNR were reported in Table 6.8. We demonstrate in Table 6.8 that using the TCN model on the feature images relatively decreased the values of

the FPR and FNR scores compared to those obtained when feature set (angles and distances) were employed.

We compared our results with (287, 326, 327) and (325) since we used the same evaluation protocol and the same metrics. However, we acknowledged the difficulty in performing a reliable comparison with other state-of-art works because of the lack of detailed pre-processing pipeline in many published works in this field. Table 6.9 illustrates our results versus the results obtained by (287, 326, 327) and (325) on the Le2i dataset and UR FD dataset, where we present different variants of our approach. We can see from this table that our results are comparable to the aforementioned state-of-the-art results. In addition, we obtained better results in terms of precision and recall which are more significant and reliable when evaluating an imbalanced dataset than accuracy and F1 score. However, our approach is independent on the background and illumination changes, unlike (287) that used optical flow and (325) which combined SVM with hidden Markov models to handle such effects. Both models depend on the RGB videos and can be influenced by illumination or occlusion.

6.4.4 Ablation study

In order to further motivate the proposed architecture, we conducted a two-stages ablation study. In the first phase, we performed a feature ablation study by removing one feature (the distance between the head and the center hip of the body). We therefore investigated the influence of this feature on the FD by comparing the results with different configurations of classifiers and data. In the second phase, we performed hyper-parameters ablation study for both SVM and LSTM classifiers employed in our model architecture. In the first phase, for the Le2i dataset, we achieved a Recall score of 100% using angles only. Similar recall score is achieved when (angle + distance) features, trained on the LSTM network, were used. In addition to the best sensitivity of 100% achieved for SVM trained on features extracted with Resnet50 from images built using angles only, the performance in terms of accuracy, precision and F1 score evaluations were quite high, in the same performance-level as that obtained with image features constructed from (angle + distance). Moreover, it is clear from Tables 6.3 and 6.10 that the results obtained from LSTM trained on angle and distance features are higher than the results obtained from angle features only in terms of accuracy, precision and F_score. Beside, the outcomes of training LSTM on both sets of features are by far better than training SVM on the same feature sets. In addition, we can see that the results obtained from the images constructed from angles and distances are better than those constructed from angles only when using either Resnet50 or Alexnet. Similar trend holds in the automatic annotation when using Resnet50 where images constructed using (angle + distance) features yielded better results than those obtained from angle-only feature images.

Table 6.10: Performance results for our FD approach on the Le2i dataset using a feature ablation study.

Features	Acc.	Precision	Recall	F_score	Spe.
Manual annotation					
Angle + SVM	0.692	0.875	0.700	0.778	0.593
Angle + LSTM	0.731	0.731	**1.000**	0.845	0.674
Feature images + AlexNet + SVM	0.769	0.818	0.900	0.857	0.786
Feature images + Resnet50 + SVM	**0.962**	**0.952**	**1.000**	**0.975**	**0.915**
Automatic annotation with pose estimation					
Angle + SVM	0.602	0.660	0.673	0.666	0.500
Angle + LSTM	0.633	0.640	0.862	0.735	0.308
Angle + TCN	0.822	0.824	0.815	0.818	0.397
Feature images + AlexNet + SVM	**0.858**	**0.888**	**0.868**	**0.878**	0.844
Feature images + Resnet50 + SVM	0.787	0.837	0.791	0.813	0.781
Feature images + TCN	0.844	0.870	0.860	0.865	**1.000**

However, the opposite trend is noticed when using Alexnet where the use of angle only feature-constructed images yields better result than (angle + distance) feature together with SVM classifier. Similarly, for the UR FD dataset, a recall score of 100% was observed (Table 6.11) when training LSTM on the set of features composed of angles only. The recall obtained for SVM classification of images extracted from our newly created images using the AlexNet activation based features was better than the one obtained from the Resnet50 activations for both sets of features (angle only, angle+distance). Moreover, from the results reported in Tables 6.6 and 6.11, we can see that exploiting distances in addition to angles increased the performance of our results on all datasets as compared to the case when angles alone were used. Again, the TCN model yielded better performance than the SVM and LSTM for the set of features composed of angles only for both datasets, except for the recall which was better using the LSTM.

In overall, the results of training LSTM and SVM on features composed of 'angle + distance' were generally better than those trained only on angles.

In the subsequent task, we performed a random search for the SVM hyper-parameters optimization where we fine-tune the most important hyper-parameters such as the kernel type and penalty value. The kernel types we considered are: linear, polynomial, RBF and sigmoid. Whereas, a log scale in the range of [10, 1.0, 0.1, 0.001] were considered for penalty value parameter. We observed that the best recall and precision values were obtained for a polynomial kernel and a penalty value of 1.0. Figure. 6.8 illustrates recall and precision values for different kernel types and penalty values for the Le2i dataset and feature images inferred from Resnet50 network.

Similarly, we performed a hyper-parameters ablation study for the LSTM model. We considered for this study the number of epochs, number of neurons and the batch size,

Table 6.11: Performance results for our FD approach on the UR FD dataset using a feature ablation study.

Features	Acc.	Precision	Recall	F_score	Spe.
Manual annotation					
Angle + SVM	0.700	0.727	0.727	0.727	0.654
Angle + LSTM	0.600	0.579	**1.000**	0.734	0.632
Feature images + AlexNet + SVM	0.950	**1.000**	0.917	**0.957**	0.789
Feature images + Resnet50 + SVM	**0.960**	**1.000**	0.833	0.907	**0.876**
Automatic annotation with pose estimation					
Angle + SVM	0.600	0.667	0.638	0.652	0.545
Angle + LSTM	0.730	0.704	0.950	0.809	0.400
Angle + TCN	0.950	0.971	0.900	0.934	0.552
Feature images + AlexNet + SVM	0.929	0.930	0.952	0.941	0.893
Feature images + Resnet50 + SVM	**0.971**	**0.976**	**0.976**	**0.976**	**0.964**
Feature images + TCN	0.700	0.744	0.779	0.761	0.916

Table 6.12: Performance results for the feature ablation study on the cross dataset 1 and its reciprocal (cross dataset 2) with pose estimation for automatic annotation.

Features	Acc.	Precision	Recall	F_score	Spe.
Cross dataset 1					
Angle + LSTM	0.670	0.671	**0.833**	0.788	0.350
Angle + TCN	0.660	0.819	0.575	0.520	0.457
Feature images + Resnet50 + SVM	**0.815**	**0.850**	0.831	**0.840**	**0.791**
Feature images + TCN	0.710	0.698	0.696	0.697	0.596
Cross dataset 2					
Angle + LSTM	0.624	0.631	0.869	0.731	0.280
Angle + TCN	0.654	0.797	0.647	0.714	0.354
Feature images + Resnet50 + SVM	**0.970**	**0.983**	**0.967**	**0.975**	**0.975**
Feature images + TCN	0.801	0.754	0.724	0.739	0.924

where their respective ranges are (10,50,100); (600,1000) and (32,64,128), respectively. Best results of recall and precision were found for a batch size of 64, 10 epochs and 600 neurons as illustrated in Fig. 6.9 for the UR FD dataset and features extracted after automatic annotation.

6.4.5 Discussion

Our prediction model yields higher recall values than specificity values in most cases, which attests that our approach allowed us to better classify positive cases over negative cases. From another point of view, a false positive may be followed with a needless action, while a false negative would not get the necessary attention (as shown in Figs. 6.10b and 6.10c, respectively), which can potentially yield dangerous situations. Also, precision was in most cases higher than recall evaluation, suggesting a higher number of false negatives than false positives. Typically, false alarms were detected when an actor bends down and then lies

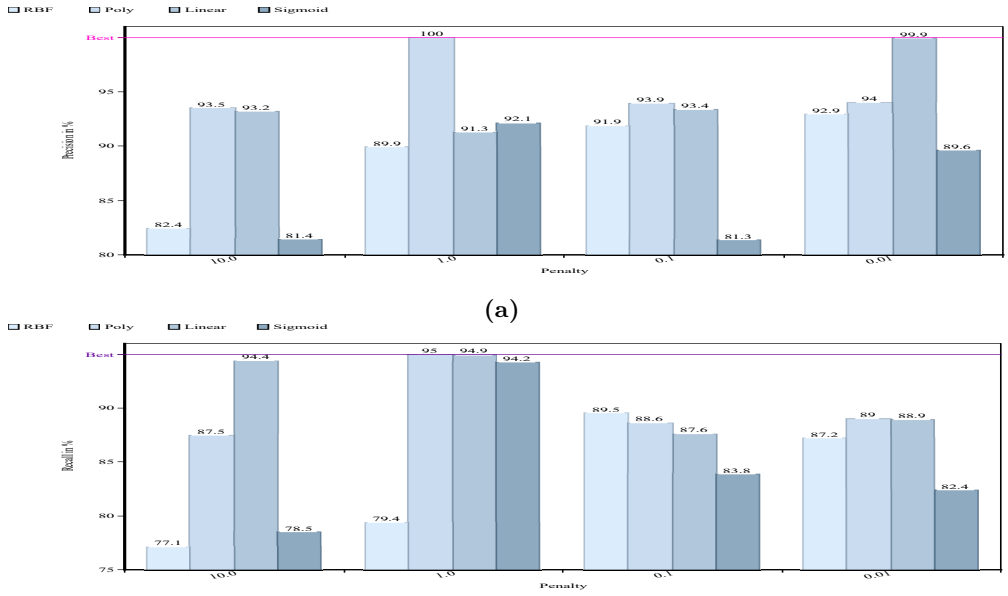


Figure 6.8: Optimization of our SVM hyper-parameters using Random search. a) represents optimization of the precision whereas b) the recall.

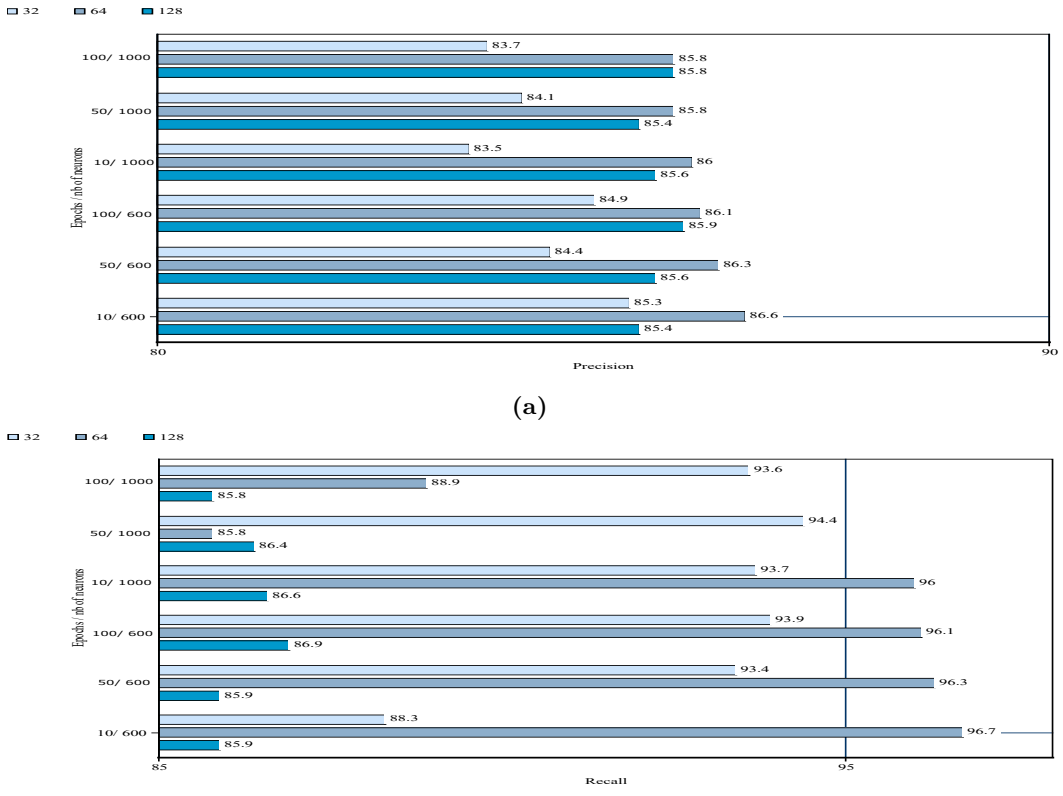


Figure 6.9: Optimization of our LSTM hyper-parameters using ablation. a) represents optimization of the precision whereas b) the recall.

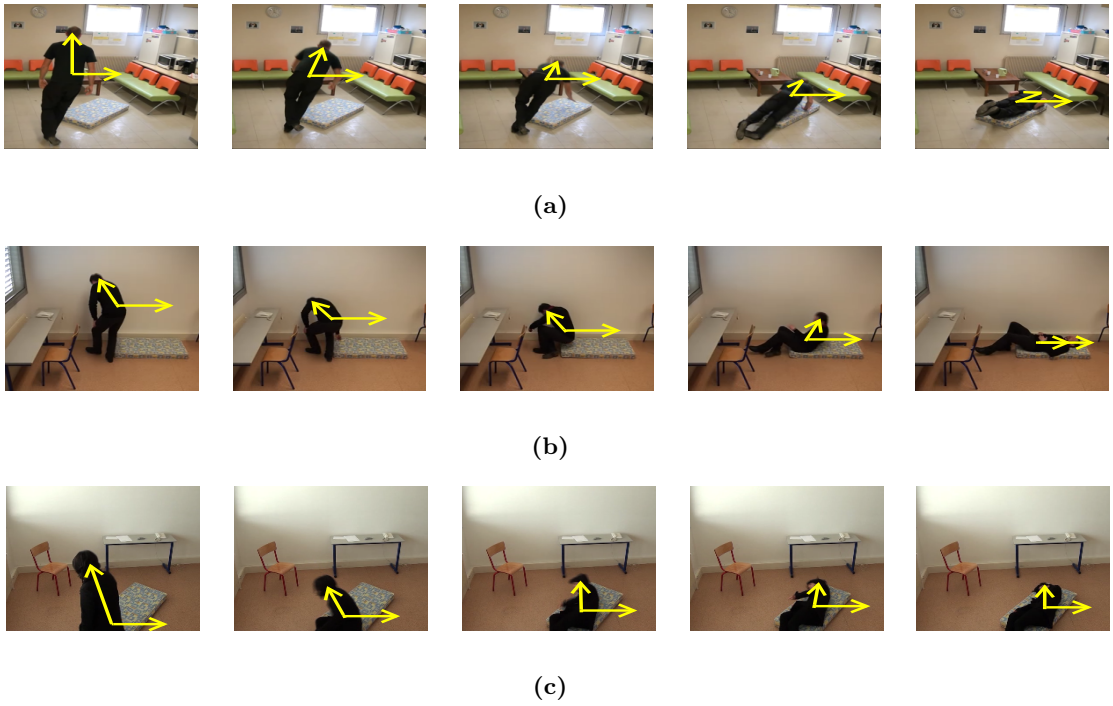


Figure 6.10: Samples from the Le2i FD dataset representing (a) Changes of the angle values across frames for a falling posture (b) False positive situation where a lying down posture is detected as a fall (c) False negative situation where a falling posture is miss detected.

down on the floor or bends down to grab something. We can see from Fig. 6.10a that a fall event could be detected by observing the sequence of angles defined above across frames. The angle value decreases to reach 0° or increases to reach 180° when a person is falling. However, Fig. 6.10b illustrates a false positive situation where the lying posture was detected as a fall since the angle value is getting closer to 0° which may lead to triggering an alarm for a 'no fall' event. Fig. 6.10c demonstrates a false negative situation where the fall was miss-detected since the value of the angle was close to 90° even after the fall occurred. This kind of situation may put the falling person in danger and deprive him from getting accurate assistance.

Vision-based FD, as a particular domain in the vision-based human activity recognition area could also be influenced by the various issues that may affect the effectiveness of these systems. Beddiar et al. (30), and Ezatzadeh et al. (328) presented an overview of many limitations and challenges that we have to cope with to provide efficient and reliable human activity recognition and FD systems. Especially, occlusion and overlapping of different people present in the monitored environment are among the major problems to FD systems, since different objects or people may be placed between the camera and the subject. To overcome this problem, researchers proposed to use multiple cameras; although, this requires to put forward mechanisms for reconciling the various views of the same subject. The field view of

the camera is another important issue in FD, because the subject moves around and could be lost by the current camera view, and if a fall occurred at that stage, then it will likely be miss-detected. Furthermore, people do not like to be watched or recorded, so privacy and intimacy issues are among other handicaps for use of such systems. Other challenges specific to FD are discussed in (328). Gathering a large realistic, multi-view dataset is still the main challenge which is not yet possible due to privacy issues and rarity of occurrence of falls in monitored scenes. Therefore, falls and daily living activities recorded in the mostly used datasets were simulated by healthy adults and not by the elderly or patients at risk of falling. Likewise many other state-of-the-art studies, this is a crucial limitation of our current work since simulated data do not accurately reflect real fall situations. For instance, geometry of the body could be altered when trying to simulate a fall, which may affect the performance of our proposed FD algorithm. Besides, our work and most of the current systems are not able to distinguish between fall, sudden sitting down, lying or crouching down as in sport activities for instance. From a computational perspective, to be efficient enough, FD systems should reduce the time it takes to alert caregivers to provide immediate support to the elderly. So minimizing computational complexity when developing real-time systems is required to avoid serious consequences of falling. Another aspect of interest to rehabilitation community but not considered by our study is related to the type of fall that has been identified. For instance, Putra et al. (329) divided falls into a broader set of categories; namely, forward, backward, left-side, right-side, blinded-forward, and blinded-backward. The direction the individual takes whilst falling, the duration of the fall, prior activities to the fall as well as the age and physical conditions of the individual are all other important aspects to devise appropriate rehabilitation strategy. Strictly speaking, our method was not designed to tackle such rehabilitation challenges, but our algorithms could provide a baseline to develop new methods to get better performance in future developments. This also calls for appropriate large scale dataset if one wants to take rehabilitation purpose in mind.

6.5 Conclusion and Future Directions

We presented in this paper an effective vision-based approach for FD based on new macro features and machine learning based methodology which extends our previous contribution (35). Our approach allowed us to construct RGB images of calculated angles and distances between the head, the center hip of the target subjects and the horizontal axis passing through the center hip. Both manual and automatic annotations of the geometrical locations of the head and the center hip of the body were contrasted. The automatic annotation makes use of pose estimation algorithm. These constructed sets of (macro) images constitute our distinctive features for the FD task. Next, SVM, TCN and LSTM classifiers were used along

with a pre-trained model to classify the created images into falls and daily life activities. We also compared the features extracted using both the Resnet50 and the Alexnet models. For the testing purpose, we used the Le2i dataset and the UR FD dataset to evaluate our approach's performance utilizing the accuracy, precision, recall, specificity and F_score evaluation metrics. Experimental results showed that the performances of our proposed approach are comparable to that of the state-of-the-art FD methods. However, some limitations are also noticed. For instance, it will be desirable to improve the approach to distinguish between lying and falling postures. Besides, in the future, we would also like to improve the automatic annotations of the head and the body center hip positioning of the individuals from video sequences. On the other hand, there is a room for improvement in the training pipeline through a better selection of training samples inputted to our SVM, TCN and LSTM classifiers and better optimization of the LSTM parameters. For instance, we have noticed the prospect of performing a cross-view evaluation to investigate the method's performance. Also, it would be interesting to explore an extensive real-world fall database that could provide a realistic understanding of the fall process for evaluating FD performance. A complete understanding of neuro-psychological factors related to the risk of falling and the relationships between them will undoubtedly help researchers compile a more comprehensive profile of individuals at high risk of falling. Furthermore, trying to predict fall before its occurrence could have significant applications. One trivial use is to trigger a fall alarm to enable caregiver to provide timely help to the victim. It can also allow us to initiate some protection mechanisms to decrease the impact of the fall injuries. For that, keeping tracks of the subject's biological parameters and its fall history by the FD system may be beneficial.

7

Summary

In this thesis, we presented two main methods for vision-based human action recognition and fall detection for elderly monitoring. Our proposals were evaluated using different benchmark datasets and experiments demonstrated that our results outperformed the state-of-the-art methods. In this chapter, we conclude our work by pointing out key contributions, research methodology, validation methodology and software/hardware tools. We discuss at the end, the limitations to our methods as well as some perspectives and future works.

In general, this thesis could mainly be partitioned into two parts: The first part includes an overview on vision-based human activity recognition and a related literature review. It aimed at introducing, investigating and analysing in depth the HAR domain. In a first chapter, we deeply explored the HAR related concepts before reviewing and focusing in the second chapter on the existing vision-based HAR techniques. This part allowed us to better understand the current progress in the HAR field and provided us with a glimpse of the strengths and limitations of existing methods of the literature. A last chapter of the first part was devoted to review the fall detection which consists in a particular area of the HAR domain. This chapter enabled us to extract the limitations related to fall detection systems and elaborate our contribution in this field. The second part, in the light of previous analysis, includes our proposed methods to resolve some of the HAR and fall detection challenges. In the first chapter, we presented a multi-modal human activity recognition solution based on deep learning to resolve the challenge of combining multiple modalities to recognize human actions. Moreover, in the second chapter, we designed a vision-based fall detection approach based on the body geometry and pose estimation to localize the human body and identify the performed actions. For both proposals, we evaluated the performance using standard benchmark datasets to identify their strengths and limitations. In the following, we discuss the key contributions related to this thesis.

7.1 Key Contributions

In this section, we discuss the key contributions of this thesis and we demonstrate that objectives listed in Chapter. 1 are realized.

- A comprehensive review of the state-of-the-art techniques of vision-based human activity recognition was proposed to review and summarize the progress of HAR systems from the computer vision perspective. The survey (30) aims to provide the reader with an up to date analysis of the literature related to vision-based HAR and recent progress in the field. At the same time, it highlights the main challenges and future directions and helps us to identify the gaps for new contributions. During this study, it is observed that almost all approaches still suffer from certain limitations. However, we noticed that deep learning-based approaches are getting more attention nowadays due to the progress they have made and the promising results in terms of detection and recognition performance. On the other hand, interactions and group activities recognition are among prominent research topics since they can provide useful information in many HAR application fields such as video surveillance, public security, abnormal activity detection, …etc. The proposed survey aims as well to realize goals 1 and 2 mentioned in chapter 1.

- Abnormal human activity recognition methods were not well covered in the literature and methods of normal human action recognition were adapted and used. This made the recognition process inefficient and the results were not really satisfying. Therefore, we provided a comprehensive overview of the state-of-the-art techniques of vision-based abnormal human activity recognition to identify limitations and challenges of recognition of such activities (4, 31). This helps us to realize goal 3 and to conduct research on abnormal human action recognition in general and FD in particular. Chapter 3 reviews some existing fall detection methods of the literature.

- One of the major challenges of human action recognition is related to the fusion of multiple modalities. Many techniques in the literature proposed to fuse two modalities such as RGB with depth or depth with skeleton data but did not mention the fusion of the three modalities. To tackle this problem, we presented in chapter 5 a multi-modal framework for human action recognition by combining RGB, Depth and skeleton data using canonical correlation analysis as a feature fusion strategy. This contribution (32) based on a supervised deep learning and a transfer learning technique from pretrained models, aimed to provide a new representation of human action in video sequences and helped us to get higher recognition accuracy by realizing objectives 4, 5 and 6. We

observed that the accuracy was improved by combining the features from each two sets of images over that using a single modality alone. Fusing the three modalities had as well improved the recognition accuracy over that using single modalities or pairwise modalities. Moreover, fusing dynamic depth and skeleton images achieved best results as they present complementary temporal features.

- Fall detection for elderly monitoring has become a very important research field for health care. It helps to identify or predict the occurrence of falls to improve the elderly quality of life, enable efficient medical assistance, reduce the falls related damages and provide them with daily health care. However, the available FD datasets contain simulated data and are very small. This presents a very challenging issue to using deep learning based approaches for this task. To overcome this limitation, and benefit from performances of the deep learning models, we adopted a transfer learning approach from pre-trained CNN models. We presented in chapter 6 an efficient method for elderly fall detection based on a supervised approach. Our proposal (35, 36, 37) relies on the body geometry and pose estimation to detect the subject, track him and identify occurring falls using angles and distances between the head and the center hip of the body. We achieved high recognition accuracy on features extracted from our newly created images using Resnet50 pretrained model and we outperformed the state-of-the-art results by significant margin when evaluating on same benchmark datasets. The goal 7 was fulfilled through this contribution.

- We compared our proposals in chapters 5 and 6 to existing vision-based human action recognition and fall detection methods respectively using standard benchmark datasets such as UTD-MHAD, NTU RGB+D and UR FD. The results outperformed the state-of-the-art results and achieved higher performances for both proposals as shown in Table 7.1, Table 7.2, Table 7.3 and Table 7.4 and this realizes our goal 8 mentioned in chapter 1.

7.2 Research Methodology

To achieve the set targets of our thesis, we followed a coherent methodology starting from identifying limitations of existing approaches to suggesting new techniques and solutions.

In a first step, we analyzed the existing HAR methods and drew a mind map of general aspects of HAR. This helped us to summarize and categorize the underlying field into main criteria and conducted us to suggest our survey paper (30). Furthermore, this analysis served us to suggest new contributions which may resolve some of the limitations and challenges

Table 7.1: Comparison of our proposed multi-modal HAR method with previous methods on the UTD-MHAD Dataset.

Method	Accuracy %
Decision Fusion Using LOGP (314)	88.40
Depth + inertial data fusion + CRC classifier (313)	79.10
5-CNN fusion of skeleton images (315)	95.38
fusion with CCA and KELM (316)	97.91
DI RGB + DI Depth + Skeleton images + LSTM (Ours)	**98.88**

Table 7.2: Comparison of the proposed multi-modal HAR method with previous methods on the NTU RGB+D Dataset.

Method	Accuracy %
Deep RNN (77)	64.09%
Deep LSTM (77)	67.29%
Joint trajectory maps + CNN (317)	75.20%
Part-aware LSTM (77)	70.20%
DI RGB + DI Depth + Skeleton images + LSTM (Ours)	**75.50%**

Table 7.3: Performance comparison of our FD approach results with other existing approaches on the Le2i dataset

Approaches	**Precision**	**Recall**	**F_score**
Combined curvlets + HMM (325)	-	98.00%	-
OF + CNN (287)	-	93.60%	-
ours: **Angle + Distance + Resnet50 + SVM**	**100%**	95.00%	**97.40%**
ours: **Angle + AlexNet + SVM**	81.80%	90.00%	85.70%
ours: **Angle + Distance + LSTM**	76.90%	**100%**	87.00%

Table 7.4: Performance comparison of our FD approach results with other existing approaches on the UR Fall detection dataset

Approaches	**Precision**	**Recall**	**F_score**
Combined curvlets + HMM (325)	-	-	-
OF + CNN (287)	-	**100%**	-
ours: **Angle + Distance + Resnet50 + SVM**	**100%**	90.00%	94.70%
ours: **Angle + AlexNet + SVM**	**100%**	91.70%	**95.70%**
ours: **Angle + Distance + LSTM**	80.00%	**100%**	88.90%

encountered by the researchers of the HAR field, among which combining different data modalities to classify various actions which is still demanding. Therefore, the multi modality fusion challenge was inspected through our paper on Multi-Modal Human Activity Recognition using deep learning (32), as shown in chapter 5. For that, we proposed a technique based on deep learning and transfer learning where we merged RGB, depth and skeleton data using canonical correlation analysis as a feature fusion strategy. Our results were satisfying and outperformed the state-of-the-art results on public benchmark datasets.

Afterwards, we scrutinized the abnormal human action detection area which was not sufficiently covered in the literature as it compromises various challenges. For instance, we examined the fall detection that consists in a particular area of abnormal human action recognition and reviewed some existing techniques in papers (4, 31, 35, 36, 37). Many applications related to health care and elderly monitoring were studied to identify the motivations behind FD systems. This conducted us to suggest new contributions in the field of FD for elderly monitoring as shown in chapter 6. Our technique uses body geometry and pose estimation and is based on deep learning to recognize falls among daily life activities. Again, our results achieved higher recognition accuracy and were comparable to the state-of-the-art results.

7.3 Validation methodology and software/hardware tools

The validation of an approach is a vital step as we mentioned in chapter 2. It is used to confirm that the analytical procedure employed for a specific test is suitable for its intended use. It has for aim to assess the correctness, the quality, the reliability and the consistency of the proposed approach by demonstrating that the results are conforming and satisfying the requirements specified in the methodology. In this thesis, the experimental methodology is used to evaluate and validate the performance of our proposed methods. We used different publicly available benchmark datasets for human action recognition and fall detection. In this regard, to validate the multi-modal human activity recognition proposal, we used the UTD-MHAD and the NTU-RGB+D datasets and to evaluate the fall detection proposal, we used the Le2i and the UR FD datasets.

- The UTD-MHAD dataset was collected using a Microsoft Kinect sensor and a wearable inertial sensor in an indoor environment. It consists of 27 simple actions performed by 8 subjects four times. The dataset includes a total of 861 data sequences and is composed of four data modalities: RGB videos, depth videos, skeleton joint positions and inertial sensor signals. For data synchronization purposes, a time stamp for each sample was recorded.

- The NTU RGB+D dataset was collected using three Kinect V2 cameras concurrently. It contains 56,880 video samples of 60 action classes performed by 40 subjects. Highly variant camera settings were used to capture four data modalities in three major categories: daily actions, mutual actions, and medical conditions.

- The Le2i fall detection dataset was collected using a single fixed camera and illustrates many difficulties of realistic video sequences of an elderly home or office, such as variable illumination and occlusion. It contains 221 videos of 131 falls, and 90 daily life activities (ADL) simulated by several actors in four different locations.

- The UR FD dataset was recorded using two Microsoft Kinect cameras from two different perspectives for fall events and from one single perspective for the ADL. It consists of 70 video sequences of 30 falls and 40 ADLs simulated by different actors.

To implement our algorithms, we used Matlab R2019 and R2020 as base software inline with python 3.6 and Keras as backend for the deep learning models. Furthermore, we used an Intel Core i5.3570 CPU with 8GB RAM as hardware and a Google Colaboratory platform with a hosted runtime and a GPU (Cuda).

7.4 Limitations

Thought our proposed methods in this thesis yielded high performances, they still have some limitations. We present in this section, some of the limitations and unresolved challenges of our proposals.

In our survey (30), we presented an overview of many limitations and challenges that we have to cope with to provide efficient and reliable human activity recognition. Despite the promising results of our multi-modal HAR method, it has not completely resolved these challenges, but our algorithms could provide a baseline to develop new methods to get better performance in future developments. Our multi-modal method is computationally expensive, and may not be suitable for real-time applications.

Vision-based fall detection could be affected by several issues that may affect the effectiveness of these systems. Moreover, since fall detection systems are mainly implemented for monitoring and assistance applications in elderly homes or offices and care centers, occlusion and overlapping of different people and objects present in the monitored environment are among the major problems. In our method, we suggest to calculate angles and distances between the head and the center hip of the subject body among the video frames. So, if objects or people are placed between the monitored subject and the camera leading to occlusion, this may influence on the features values. Moreover, The field view is another limitation to our

proposal because we may lose the subject from the camera when he moves around and if fall happens in this hidden area, it will be miss-detected. Using multiple cameras can overcome these two problems, but is also challenging for our system due to issues related to synchronisation and reconstruction of the scene from different views. Another important issue is related to privacy and intimacy because people do not like to be watched or recorded, which may interrupt the use of our proposal. Besides, simulated data and lack of large benchmark datasets is still an open challenge not only to our proposal but to the majority of FD systems. Simulated data was recorded by healthy adults and do not accurately reflect real falls and ADLs. Furthermore, current systems are not able yet to distinguish between fall, sudden sitting down, lying or crouching down that is also the case of our proposal.

As a conclusion, we can say that our developed methodologies are still facing many problems and require many more research. However, we are very optimistic about the future works since some issues can be solved soon.

7.5 Future works

As the reader may have noticed through this thesis, we presented in this thesis two main methods based on machine learning and deep learning algorithms for HAR and FD. Our algorithms are not specific to a particular field but could be generalized to different contexts.

To enhance the performance of our HAR proposed methodology, it is possible to explore more fusion strategies at different levels; raw data fusion, feature fusion and decision fusion. Some data augmentation methods could also be contemplated to enrich the datasets, which may help to increase the reliability of the deep learning algorithm used for activities classification. Besides, we believe there is also a room for further improvement on the recognition accuracy achieved by NTU RGB+D dataset throughout a more fine-gained optimization of the parameters of the underlined LSTM model. Extension on multiple views could also augment the efficiency of our proposal and widen its application field. Besides, transfer learning was partly used for feature extraction, we wish to extend this for a full transfer learning technique.

On the other hand, there is a room for improvement of our FD methodology in the training pipeline through a better selection of training samples inputted to our SVM and LSTM classifiers and better-optimizing the LSTM parameters. For instance, we have noticed the prospect of performing a cross-view evaluation to investigate the method's performance when different perspectives are studied. Also, it would be interesting to explore an extensive real-world fall database that could provide an enhanced understanding of the fall process for evaluating fall detection performance. A complete understanding of neuropsychological factors related to the risk of falling and the relationships between them will undoubtedly

help researchers compile a more comprehensive profile of individuals at high risk of falling. Furthermore, trying to predict fall before its occurrence could have significant applications. One trivial use is to trigger a fall alarm to provide immediate help to the victim by the caregivers. It can also allow us to initiate some protection mechanisms to decrease the impact of the fall injuries. For that keeping tracks of the subject's biological parameters and its fall history by the FD system may be beneficial. Besides, in the future, we would also like to work on ensuring the privacy and intimacy of the monitored subjects by omitting the RGB images.

References

[1] Daniel Weinland, Remi Ronfard, and Edmond Boyer. **A survey of vision-based methods for action representation, segmentation and recognition**. *Computer vision and image understanding*, **115**(2):224–241, 2011.

[2] Uzair Asad Akansha, Mishra Shailendra, and Narayan Singh. **Analytical review on video-based human activity recognition**. In *Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on*, pages 3839–3844. IEEE, 2016.

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

[4] Djamila Romaissa Beddiar and Brahim Nini. **Vision based abnormal human activities recognition: An overview**. *2017 8th International Conference on Information Technology (ICIT)*, pages 548–553, 2017.

[5] Aashni Haria, Archanasri Subramanian, Nivedhitha Asokkumar, Shristi Poddar, and Jyothi S Nayak. **Hand gesture recognition for human computer interaction**. *Procedia Computer Science*, **115**:367–374, 2017.

[6] K Martin Sagayam and D Jude Hemanth. **Hand posture and gesture recognition techniques for virtual reality applications: a survey**. *Virtual Reality*, **21**(2):91–107, 2017.

[7] Jing Shao, Kai Kang, Chen Change Loy, and Xiaogang Wang. **Deeply learned attributes for crowded scene understanding**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4657–4666, 2015.

[8] Wenkai Xu and Eung-Joo Lee. **A Novel Method for Hand Posture Recognition Based on Depth Information Descriptor.** *KSII Transactions on Internet & Information Systems*, **9**(2), 2015.

[9] Adam Poulos, Cameron Brown, Daniel McCulloch, and Jeff Cole. **Context-aware augmented reality object commands**, October 17 2017. US Patent 9,791,921.

[10] Siddharth S. Rautaray and Anupam Agrawal. **Vision Based Hand Gesture Recognition for Human Computer Interaction: A Survey**. *Artif. Intell. Rev.*, **43**(1):1–54, January 2015.

[11] Svitlana Antoshchuk, Mykyta Kovalenko, and Jürgen Sieck. **Gesture Recognition-Based Human–Computer Interaction Interface for Multimedia Applications**. In *Digitisation of Culture: Namibian and International Perspectives*, pages 269–286. Springer, 2018.

[12] Mohammad Sabokrou, Masoud Pourreza, Mohsen Fayyaz, Rahim Entezari, Mahmood Fathy, Jürgen Gall, and Ehsan Adeli. **AVID: Adversarial Visual Irregularity Detection**. In *Asian Computer Vision Conference*, 2018.

[13] Mohammad Sabokrou, Mohammad Khalooei, and Ehsan Adeli. **Self-Supervised Representation Learning via Neighborhood-Relational Encoding**. *International Conference on Computer Vision*, 2019.

[14] Mohammad Sabokrou, Mahmood Fathy, Zahra Moayed, and Reinhard Klette. **Fast and accurate detection and localization of abnormal behavior in crowded scenes**. *Machine Vision and Applications*, **28**(8):965–985, 2017.

[15] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, Zahra Moayed, and Reinhard Klette. **Deep-anomaly: Fully convolutional neural network for fast anomaly detection in crowded scenes**. *Computer Vision and Image Understanding*, **172**:88–97, 2018.

[16] Mohammad Sabokrou, Mahmood Fathy, Mojtaba Hoseini, and Reinhard Klette. **Real-time anomaly detection and localization in crowded scenes**. In *Proceedings of the IEEE CVPR Workshops*, pages 56–62, 2015.

[17] M Sabokrou, M Fathy, and M Hoseini. **Video anomaly detection and localisation based on the sparsity and reconstruction error of auto-encoder**. *Electronics Letters*, **52**(13):1122–1124, 2016.

[18] Mohammad Sabokrou, Mohsen Fayyaz, Mahmood Fathy, and Reinhard Klette. **Deep-cascade: Cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes**. *IEEE Transactions on Image Processing*, **26**(4):1992–2004, 2017.

[19] Allah Bux. *Vision-based human action recognition using machine learning techniques*. PhD thesis, Lancaster University, 2017.

[20] Guodong Guo and Alice Lai. **A survey on still image based human action recognition**. *Pattern Recognition*, **47**(10):3343–3361, 2014.

[21] Manoj Ramanathan, Wei-Yun Yau, and Eam Khwang Teoh. **Human action recognition with video data: research and evaluation challenges**. *IEEE Transactions on human-machine systems*, **44**(5):650–663, 2014.

[22] Guangchun Cheng, Yiwen Wan, Abdullah N Saudagar, Kamesh Namuduri, and Bill P Buckles. **Advances in human action recognition: A survey**. *arXiv preprint arXiv:1501.05964*, 2015.

[23] Roanna Lun and Wenbing Zhao. **A survey of applications and human motion recognition with microsoft kinect**. *International Journal of Pattern Recognition and Artificial Intelligence*, **29**(05):1555008, 2015.

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

[25] Maryam Ziaeefard and Robert Bergevin. **Semantic human activity recognition: a literature review**. *Pattern Recognition*, **48**(8):2329–2345, 2015.

[26] Eisa Jafari Amirbandi and Ghazal Shamsipour. **Exploring methods and systems for vision based human activity recognition**. In *Swarm Intelligence and Evolutionary Computation (CSIEC), 2016 1st Conference on*, pages 160–164. IEEE, 2016.

[27] Soo Min Kang and Richard P Wildes. **Review of action recognition and detection methods**. *arXiv preprint arXiv:1610.06906*, 2016.

[28] T Subetha and S Chitrakala. **A Survey on human activity recognition from videos**. In *Information Communication and Embedded Systems (ICICES), 2016 International Conference on*, pages 1–7. IEEE, 2016.

REFERENCES

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

[30] Djamila Romaissa Beddiar, Brahim Nini, Mohammad Sabokrou, and Abdenour Hadid. **Vision-based human activity recognition: a survey**. *Multimedia Tools and Applications*, **79**(41):30509–30555, 2020.

[31] Djamila Romaissa Beddiar and Brahim Nini. **Abnormal human activities recognition: brief synthesis of vision based fall detection**.

[32] Beddiar Djamila Romaissa, Oussalah Mourad, and Nini Brahim. **Vision-Based Multi-Modal Framework for Action Recognition**. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5859–5866. IEEE, 2021.

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

[34] Basura Fernando, Efstratios Gavves, Jose M Oramas, Amir Ghodrati, and Tinne Tuytelaars. **Modeling video evolution for action recognition**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5378–5387, 2015.

[35] Beddiar Djamila Romaissa, Oussalah Mourad, Nini Brahim, and Bounab Yazid. **Fall Detection using Body Geometry in Video Sequences**. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–5. IEEE, 2020.

[36] Beddiar Djamila Romaissa, Oussalah Mourad, Nini Brahim, and Bounab Yazid. **Vision-Based Fall Detection Using Body Geometry**. In *International Conference on Pattern Recognition*, pages 170–185. Springer, 2021.

[37] Djamila Romaissa Beddiar, Mourad Oussalah, and Brahim Nini. **Fall detection using body geometry and human pose estimation in video sequences**. *Journal of Visual Communication and Image Representation*, page 103407, 2021.

[38] Tingting Liu, Zengzhao Chen, Hai Liu, Zhaoli Zhang, and Yingying Chen. **Multi-modal Hand Gesture Designing in Multi-screen Touchable Teaching System for Human-computer Interaction**. In *Proceedings of the 2Nd International Conference on Advances in Image Processing*, ICAIP '18, pages 198–202, New York, NY, USA, 2018. ACM.

[39] Nadipuram R Prasad, Jason C King, and Thomas Lu. **Machine intelligence-based decision-making (MIND) for automatic anomaly detection**. In *Optical Pattern Recognition XVIII*, **6574**, page 65740F. International Society for Optics and Photonics, 2007.

[40] Derek Hao Hu, Xian-Xing Zhang, Jie Yin, Vincent Wenchen Zheng, and Qiang Yang. **Abnormal activity recognition based on hdp-hmm models.** In *IJCAI*, pages 1715–1720, 2009.

[41] Marina L Gavrilova, Yingxu Wang, Faisal Ahmed, and Padma Polash Paul. **Kinect Sensor Gesture and Activity Recognition: New Applications for Consumer Cognitive Systems**. *IEEE Consumer Electronics Magazine*, **7**(1):88–94, 2018.

[42] Kaiping Xu, Zheng Qin, and Guolong Wang. **Recognize human activities from multi-part missing videos**. In *IEEE International Conference on Multimedia and Expo, ICME 2016, Seattle, WA, USA, July 11-15, 2016*, pages 1–6, 2016.

[43] Mohamed Ben Youssef, Imen Trabelsi, and Med Bouhlel. **Human Action Analysis for Assistance with Daily Activities**. *International Journal on Human Machine Interaction*, 07 2016.

[44] Jessica PM Vital, Diego R Faria, Gonçalo Dias, Micael S Couceiro, Fernanda Coutinho, and Nuno MF Ferreira. **Combining discriminative spatiotemporal features for daily life activity recognition using wearable motion sensing suit**. *Pattern Analysis and Applications*, **20**(4):1179–1194, 2017.

[45] Pedro Chahuara, Anthony Fleury, Michel Vacher, and François Portet. **Méthodes SVM et MLN pour la reconnaissance automatique d'activités humaines dans les habitats perceptifs: tests et perspectives**. In *RFIA 2012 (Reconnaissance des Formes et Intelligence Artificielle)*, pages 978–2–9539515–2–3, Lyon, France, January 2012. Session "Posters".

[46] Rashim Bhardwaj and Pradeep Kumar Singh. **Analytical review on human activity recognition in video**. In *Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference*, pages 531–536. IEEE, 2016.

[47] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa. **Human Action Recognition by Representing 3D Skeletons As Points in a Lie Group**. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 588–595, Washington, DC, USA, 2014. IEEE Computer Society.

[48] Helena Carolina Teixeira Lopes. **Contextual game design: from interface development to human activity recognition**. 2017.

[49] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. **Going deeper into action recognition: A survey**. *Image and vision computing*, **60**:4–21, 2017.

[50] GRS Murthy and RS Jadon. **A review of vision based hand gestures recognition**. *International Journal of Information Technology and Knowledge Management*, **2**(2):405–410, 2009.

[51] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. **Real-time human pose recognition in parts from single depth images**. *Communications of the ACM*, **56**(1):116–124, 2013.

[52] Saba Jadooki, Dzulkifli Mohamad, Tanzila Saba, Abdulaziz S Almazyad, and Amjad Rehman. **Fused features mining for depth-based hand gesture recognition to classify blind human communication**. *Neural Computing and Applications*, **28**(11):3285–3294, 2017.

[53] Nhan Nguyen-Duc-Thanh, Daniel Stonier, SungYoung Lee, and Dong-Han Kim. **A new approach for human-robot interaction using human body language**. In *International Conference on Hybrid Information Technology*, pages 762–769. Springer, 2011.

[54] Xu Wenkai and Eung-Joo Lee. **Continuous gesture trajectory recognition system based on computer vision**. *International Journal of Applied Mathematics and Information Sciences*, pages 339–346, 2012.

[55] Nicolas Mollet and Ryad Chellali. **Détection ET interprétation des Gestes de la Main**. In *2005 3rd International Conference on SETIT, year=2005*.

[56] Tianmin Shu, Dan Xie, Brandon Rothrock, Sinisa Todorovic, and Song Chun Zhu. **Joint inference of groups, events and human roles in aerial videos**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4576–4584, 2015.

REFERENCES

[57] Michael S Ryoo and Jake K Aggarwal. **Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities**. In *Computer vision, 2009 ieee 12th international conference on*, pages 1593–1600. IEEE, 2009.

[58] Kishore K Reddy and Mubarak Shah. **Recognizing 50 human action categories of web videos**. *Machine Vision and Applications*, **24**(5):971–981, 2013.

[59] F. Negin, M. Koperski, C. F. Crispim, F. Bremond, S. Coşar, and K. Avgerinakis. **A hybrid framework for online recognition of activities of daily living in real-world settings**. In *2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 37–43, Aug 2016.

[60] Thi-Hoa-Cuc Nguyen, Jean-Christophe Nebel, Francisco Florez-Revuelta, et al. **Recognition of activities of daily living with egocentric vision: A review**. *Sensors*, **16**(1):72, 2016.

[61] Ivan Miguel Pires, Nuno M Garcia, Nuno Pombo, and Francisco Flórez-Revuelta. **From data acquisition to data fusion: a comprehensive review and a roadmap for the identification of activities of daily living using mobile devices**. *Sensors*, **16**(2):184, 2016.

[62] Hedi Tabia, Michèle Gouiffes, Lionel Lacassagne, and Bures sur Yvette. **Reconnaissance des activités humaines à partir des vecteurs de mouvement quantifiés**. 05 2012.

[63] Chuang Gan, Naiyan Wang, Yi Yang, Dit-Yan Yeung, and Alex G Hauptmann. **Devnet: A deep event network for multimedia event detection and evidence recounting**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2568–2577, 2015.

[64] Kevin Tang, Bangpeng Yao, Li Fei-Fei, and Daphne Koller. **Combining the right features for complex event recognition**. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2696–2703, 2013.

[65] Sen Wang, Zhigang Ma, Yi Yang, Xue Li, Chaoyi Pang, and Alexander G Hauptmann. **Semi-supervised multiple feature analysis for action recognition**. *IEEE Transactions on Multimedia*, **16**(2):289–298, 2014.

[66] Mihalis A. Nicolaou, Vladimir Pavlovic, and Maja Pantic. **Dynamic Probabilistic CCA for Analysis of Affective Behavior and Fusion of Continuous Annotations**. *IEEE Trans. Pattern Anal. Mach. Intell.*, **36**(7):1299–1311, July 2014.

[67] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. **Mining actionlet ensemble for action recognition with depth cameras**. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1290–1297. IEEE, 2012.

[68] Vinay Kumar, Ankur Chaturvedi, and Anjani Kumar Rai. **A Framework Using Multiple Features to Detect Multi-View Human Activity**. In *Proceedings of 3rd International Conference on Internet of Things and Connected Technologies (ICIoTCT)*, pages 26–27, 2018.

[69] Enrique Garcia-Ceja, Carlos E Galván-Tejada, and Ramon Brena. **Multi-view stacking for activity recognition with sound and accelerometer data**. *Information Fusion*, **40**:45–56, 2018.

[70] Brandon Paulson, Danielle Cummings, and Tracy Hammond. **Object interaction detection using hand posture cues in an office setting**. *International journal of human-computer studies*, **69**(1-2):19–29, 2011.

[71] François Portet, Michel Vacher, Caroline Golanski, Camille Roux, and Brigitte Meillon. **Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects**. *Personal and Ubiquitous Computing*, **17**(1):127–144, 2013.

[72] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. **Action recognition based on a bag of 3d points**. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 9–14. IEEE, 2010.

[73] Lu Xia, Chia-Chih Chen, and Jake K Aggarwal. **View invariant human action recognition using histograms of 3d joints**. In *Computer vision and pattern recognition workshops (CVPRW), 2012 IEEE computer society conference on*, pages 20–27. IEEE, 2012.

[74] Enjie Ghorbel, Rémi Boutteau, Jacques Boonaert, Xavier Savatier, and Stéphane Lecoeuche. **Kinematic Spline Curves: A temporal invariant descriptor for fast action recognition**. *Image and Vision Computing*, **77**:60–71, 2018.

[75] Simon Fothergill, Helena Mentis, Pushmeet Kohli, and Sebastian Nowozin. **Instructing people for training gestural interactive systems**. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1737–1746. ACM, 2012.

[76] Mounir Hammouche, Enjie Ghorbel, Anthony Fleury, and Sébastien Ambellouis. **Toward a real time view-invariant 3d action recognition**. In *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, 2016.

[77] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. **NTU RGB+ D: A large scale dataset for 3D human activity analysis**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016.

[78] Zequn Zhang, Yuanning Liu, Ao Li, and Minghui Wang. **A novel method for user-defined human posture recognition using Kinect**. In *Image and Signal Processing (CISP), 2014 7th International Congress on*, pages 736–740. IEEE, 2014.

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

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

[81] Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. **Human action recognition by learning bases of action attributes and parts**. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1331–1338. IEEE, 2011.

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

[83] Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, Lorenzo Seidenari, and Giuseppe Serra. **Effective codebooks for human action categorization**. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 506–513. IEEE, 2009.

REFERENCES

[84] Wenhui Li, Yongkang Wong, An-An Liu, Yang Li, Yu-Ting Su, and Mohan Kankanhalli. **Multi-Camera Action Dataset (MCAD): A Dataset for Studying Non-overlapped Cross-Camera Action Recognition**. *CoRR*, **abs/1607.06408**, 2016.

[85] Sebastian Stein and Stephen J McKenna. **Combining embedded accelerometers with computer vision for recognizing food preparation activities**. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 729–738. ACM, 2013.

[86] Mikel D Rodriguez, Javed Ahmed, and Mubarak Shah. **Action mach a spatio-temporal maximum average correlation height filter for action recognition**. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

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

[88] Anh T Nghiem, Francois Bremond, Monique Thonnat, and Valery Valentin. **ETISEO, performance evaluation for video surveillance systems**. In *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pages 476–481. IEEE, 2007.

[89] Juan Carlos Niebles, Chih-Wei Chen, and Li Fei-Fei. **Modeling temporal structure of decomposable motion segments for activity classification**. In *European conference on computer vision*, pages 392–405. Springer, 2010.

[90] Michael S Ryoo and JK Aggarwal. **UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA)**. In *IEEE International Conference on Pattern Recognition Workshops*, **2**, page 4, 2010.

[91] Chia-Chih Chen, M. S. Ryoo, and J. K. Aggarwal. **UT-Tower Dataset: Aerial View Activity Classification Challenge**. http://cvrc.ece.utexas.edu/SDHA2010/Aerial_View_Activity.html, 2010.

[92] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles. **ActivityNet: A large-scale video benchmark for human activity understanding**. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, June 2015.

[93] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. **The kinetics human action video dataset**. *arXiv preprint arXiv:1705.06950*, 2017.

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

[95] Ivan Laptev, Marcin Marszałek, Cordelia Schmid, and Benjamin Rozenfeld. **Learning Realistic Human Actions from Movies**. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2008.

[96] Marcin Marszałek, Ivan Laptev, and Cordelia Schmid. **Actions in Context**. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2009.

[97] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. **UCF101: A dataset of 101 human actions classes from videos in the wild**. *arXiv preprint arXiv:1212.0402*, 2012.

[98] Jingen Liu, Jiebo Luo, and Mubarak Shah. **Recognizing realistic actions from videos "in the wild"**. In *Computer vision and pattern recognition, 2009. CVPR 2009. IEEE conference on*, pages 1996–2003. IEEE, 2009.

[99] Giovanni Denina, Bir Bhanu, Hoang Thanh Nguyen, Chong Ding, Ahmed Kamal, Chinya Ravishankar, Amit Roy-Chowdhury, Allen Ivers, and Brenda Varda. **Videoweb dataset for multi-camera activities and non-verbal communication**. In *Distributed Video Sensor Networks*, pages 335–347. Springer, 2011.

[100] Lorenzo Seidenari, Vincenzo Varano, Stefano Berretti, Alberto Bimbo, and Pietro Pala. **Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 479–485, 2013.

[101] David Minnen, Tracy Westeyn, Thad Starner, J Ward, and Paul Lukowicz. **Performance metrics and evaluation issues for continuous activity recognition**. *Performance Metrics for Intelligent Systems*, **4**, 2006.

[102] Wenxiong Kang and Feiqi Deng. **Research on intelligent visual surveillance for public security**. In *Computer and Information Science, 2007. ICIS 2007. 6th IEEE/ACIS International Conference on*, pages 824–829. IEEE, 2007.

[103] Volker Krüger, Danica Kragic, Aleš Ude, and Christopher Geib. **The meaning of action: a review on action recognition and mapping**. *Advanced robotics*, **21**(13):1473–1501, 2007.

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

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

[106] Jake K Aggarwal and Michael S Ryoo. **Human activity analysis: A review**. *ACM Computing Surveys (CSUR)*, **43**(3):16, 2011.

[107] Zhengyou Zhang. **Microsoft kinect sensor and its effect**. *IEEE multimedia*, **19**(2):4–10, 2012.

[108] Shian-Ru Ke, Hoang Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi. **A review on video-based human activity recognition**. *Computers*, **2**(2):88–131, 2013.

[109] Sarvesh Vishwakarma and Anupam Agrawal. **A survey on activity recognition and behavior understanding in video surveillance**. *The Visual Computer*, **29**(10):983–1009, 2013.

[110] Hong-Bo Zhang, Yi-Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du, and Duan-Sheng Chen. **A comprehensive survey of vision-based human action recognition methods**. *Sensors*, **19**(5):1005, 2019.

[111] Beibei Zhan, Dorothy N Monekosso, Paolo Remagnino, Sergio A Velastin, and Li-Qun Xu. **Crowd analysis: a survey**. *Machine Vision and Applications*, **19**(5-6):345–357, 2008.

[112] Liliana Lo Presti and Marco La Cascia. **3D skeleton-based human action classification: A survey**. *Pattern Recognition*, **53**:130–147, 2016.

[113] Fei Han, Brian Reily, William Hoff, and Hao Zhang. **Space-time representation of people based on 3D skeletal data: A review**. *Computer Vision and Image Understanding*, **158**:85–105, 2017.

[114] Anshika Sharma, Pradeep Kumar Singh, and Palak Khurana. **Analytical review on object segmentation and recognition**. In *Cloud System and Big Data Engineering (Confluence), 2016 6th International Conference*, pages 524–530. IEEE, 2016.

[115] Duc Thanh Nguyen, Wanqing Li, and Philip O Ogunbona. **Human detection from images and videos: A survey**. *Pattern Recognition*, **51**:148–175, 2016.

[116] Leonardo Onofri, Paolo Soda, Mykola Pechenizkiy, and Giulio Iannello. **A survey on using domain and contextual knowledge for human activity recognition in video streams**. *Expert Systems with Applications*, **63**:97–111, 2016.

[117] Jungong Han, Ling Shao, Dong Xu, and Jamie Shotton. **Enhanced computer vision with microsoft kinect sensor: A review**. *IEEE transactions on cybernetics*, **43**(5):1318–1334, 2013.

[118] Mao Ye, Qing Zhang, Liang Wang, Jiejie Zhu, Ruigang Yang, and Juergen Gall. **A survey on human motion analysis from depth data**. In *Time-of-flight and depth imaging. sensors, algorithms, and applications*, pages 149–187. Springer, 2013.

[119] Simon Ruffieux, Denis Lalanne, Elena Mugellini, and Omar Abou Khaled. **A survey of datasets for human gesture recognition**. In *International Conference on Human-Computer Interaction*, pages 337–348. Springer, 2014.

[120] Hong Cheng, Lu Yang, and Zicheng Liu. **Survey on 3D Hand Gesture Recognition**. *IEEE Trans. Circuits Syst. Video Techn.*, **26**(9):1659–1673, 2016.

[121] Sergio Escalera, Vassilis Athitsos, and Isabelle Guyon. **Challenges in multi-modal gesture recognition**. In *Gesture Recognition*, pages 1–60. Springer, 2017.

[122] Ming Jin Cheok, Zaid Omar, and Mohamed Hisham Jaward. **A review of hand gesture and sign language recognition techniques**. *International Journal of Machine Learning and Cybernetics*, **10**(1):131–153, 2019.

[123] Liang Wang, Weiming Hu, and Tieniu Tan. **Recent developments in human motion analysis**. *Pattern recognition*, **36**(3):585–601, 2003.

[124] Weiming Hu, Tieniu Tan, Liang Wang, and Steve Maybank. **A survey on visual surveillance of object motion and behaviors**. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **34**(3):334–352, 2004.

[125] Oluwatoyin P Popoola and Kejun Wang. **Video-based abnormal human behavior recognition—A review**. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **42**(6):865–878, 2012.

[126] Paulo Vinicius Koerich Borges, Nicola Conci, and Andrea Cavallaro. **Video-based human behavior understanding: A survey**. *IEEE transactions on circuits and systems for video technology*, **23**(11):1993–2008, 2013.

[127] Natalia Díaz Rodríguez, Manuel P Cuéllar, Johan Lilius, and Miguel Delgado Calvo-Flores. **A survey on ontologies for human behavior recognition**. *ACM Computing Surveys (CSUR)*, **46**(4):43, 2014.

[128] Elias Alevizos, Anastasios Skarlatidis, Alexander Artikis, and Georgios Paliouras. **Probabilistic complex event recognition: a survey**. *ACM Computing Surveys (CSUR)*, **50**(5):71, 2017.

[129] Adriana Tapus, Antonio Bandera, Ricardo Vazquez-Martin, and Luis V Calderita. **Perceiving the person and their interactions with the others for social robotics–a review**. *Pattern Recognition Letters*, **118**:3–13, 2019.

[130] AA Afiq, MA Zakariya, MN Saad, AA Nurfarzana, Mohd Haris M Khir, AF Fadzil, A Jale, W Gunawan, ZAA Izuddin, and M Faizari. **A review on classifying abnormal behavior in crowd scene**. *Journal of Visual Communication and Image Representation*, **58**:285–303, 2019.

[131] Jason M Grant and Patrick J Flynn. **Crowd scene understanding from video: a survey**. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, **13**(2):1–23, 2017.

[132] G Sreenu and MA Saleem Durai. **Intelligent video surveillance: a review through deep learning techniques for crowd analysis**. *Journal of Big Data*, **6**(1):48, 2019.

[133] Pierre Bour, Emile Cribelier, and Vasileios Argyriou. **Crowd behavior analysis from fixed and moving cameras**. In *Multimodal Behavior Analysis in the Wild*, pages 289–322. Elsevier, 2019.

[134] Andrea Prati, Caifeng Shan, and Kevin I-Kai Wang. **Sensors, vision and networks: From video surveillance to activity recognition and health monitoring**. *Journal of ambient intelligence and smart environments*, **11**(1):5–22, 2019.

[135] Henry Friday Nweke, Ying Wah Teh, Ghulam Mujtaba, and Mohammed Ali Al-garadi. **Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions**. *Information Fusion*, **46**:147–170, 2019.

[136] Chhavi Dhiman and Dinesh Kumar Vishwakarma. **A review of state-of-the-art techniques for abnormal human activity recognition**. *Engineering Applications of Artificial Intelligence*, **77**:21–45, 2019.

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

[138] Thomas B Moeslund and Erik Granum. **A survey of computer vision-based human motion capture**. *Computer vision and image understanding*, **81**(3):231–268, 2001.

[139] David A Forsyth, Okan Arikan, Leslie Ikemoto, James O'Brien, Deva Ramanan, et al. **Computational studies of human motion: part 1, tracking and motion synthesis**. *Foundations and Trends® in Computer Graphics and Vision*, **1**(2–3):77–254, 2006.

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

[141] Lulu Chen, Hong Wei, and James Ferryman. **A survey of human motion analysis using depth imagery**. *Pattern Recognition Letters*, **34**(15):1995–2006, 2013.

[142] Yunji Liang, Xingshe Zhou, Zhiwen Yu, and Bin Guo. **Energy-efficient motion related activity recognition on mobile devices for pervasive healthcare**. *Mobile Networks and Applications*, **19**(3):303–317, 2014.

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

[144] Huy-Hieu Pham, Louahdi Khoudour, Alain Crouzil, Pablo Zegers, and Sergio Alejandro Velastin Carroza. **Video-based human action recognition using deep learning: a review**. 2015.

REFERENCES

[145] Maryam Asadi-Aghbolaghi, Albert Clapes, Marco Bellantonio, Hugo Jair Escalante, Víctor Ponce-López, Xavier Baró, Isabelle Guyon, Shohreh Kasaei, and Sergio Escalera. **A survey on deep learning based approaches for action and gesture recognition in image sequences**. In *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*, pages 476–483. IEEE, 2017.

[146] Pichao Wang, Wanqing Li, Philip Ogunbona, Jun Wan, and Sergio Escalera. **RGB-D-based human motion recognition with deep learning: A survey**. *Computer Vision and Image Understanding*, 2018.

[147] Guangle Yao, Tao Lei, and Jiandan Zhong. **A review of Convolutional-Neural-Network-based action recognition**. *Pattern Recognition Letters*, **118**:14–22, 2019.

[148] Haowei Liu, Rogerio Feris, and Ming-Ting Sun. **Benchmarking datasets for human activity recognition**. In *Visual Analysis of Humans*, pages 411–427. Springer, 2011.

[149] Jose M Chaquet, Enrique J Carmona, and Antonio Fernández-Caballero. **A survey of video datasets for human action and activity recognition**. *Computer Vision and Image Understanding*, **117**(6):633–659, 2013.

[150] Tal Hassner. **A critical review of action recognition benchmarks**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 245–250, 2013.

[151] Yulan Guo, Jun Zhang, Min Lu, Jianwei Wan, and Yanxin Ma. **Benchmark datasets for 3D computer vision**. In *Industrial Electronics and Applications (ICIEA), 2014 IEEE 9th Conference on*, pages 1846–1851. IEEE, 2014.

[152] Michael Edwards, Jingjing Deng, and Xianghua Xie. **From pose to activity: Surveying datasets and introducing CONVERSE**. *Computer Vision and Image Understanding*, **144**:73–105, 2016.

[153] Tej Singh and Dinesh Kumar Vishwakarma. **Video benchmarks of human action datasets: a review**. *Artificial Intelligence Review*, **52**(2):1107–1154, 2019.

[154] Orit Kliper-Gross, Tal Hassner, and Lior Wolf. **The action similarity labeling challenge**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **34**(3):615–621, 2012.

[155] Kai Berger. **The role of rgb-d benchmark datasets: an overview**. *arXiv preprint arXiv:1310.2053*, 2013.

[156] Jing Zhang, Wanqing Li, Philip O Ogunbona, Pichao Wang, and Chang Tang. **RGB-D-based action recognition datasets: A survey**. *Pattern Recognition*, **60**:86–105, 2016.

[157] Michael Firman. **RGBD datasets: Past, present and future**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 19–31, 2016.

[158] Tej Singh and Dinesh Kumar Vishwakarma. **Human Activity Recognition in Video Benchmarks: A Survey**. In *Advances in Signal Processing and Communication*, pages 247–259. Springer, 2019.

[159] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. **Spatio-Temporal Attention-Based LSTM Networks for 3D Action Recognition and Detection**. *IEEE Transactions on Image Processing*, **27**(7):3459–3471, 2018.

[160] Shujah Islam, Tehreem Qasim, Muhammad Yasir, Naeem Bhatti, Hasan Mahmood, and Muhammad Zia. **Single-and two-person action recognition based on silhouette shape and optical point descriptors**. *Signal, Image and Video Processing*, **12**(5):853–860, 2018.

[161] Federico Angelini, Zeyu Fu, Sergio A Velastin, Jonathon A Chambers, and Syed Mohsen Naqvi. **3D-Hog Embedding Frameworks for Single and Multi-Viewpoints Action Recognition Based on Human Silhouettes**. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4219–4223. IEEE, 2018.

[162] L Gonzalez, SA Velastin, and G Acuna. **Silhouette-based human action recognition with a multi-class support vector machine**. 2018.

[163] Unaiza Ahsan, Chen Sun, and Irfan Essa. **DiscrimNet: Semi-Supervised Action Recognition from Videos using Generative Adversarial Networks**. *arXiv preprint arXiv:1801.07230*, 2018.

[164] Earnest Paul Ijjina and Krishna Mohan Chalavadi. **Human action recognition in RGB-D videos using motion sequence information and deep learning**. *Pattern Recognition*, **72**:504–516, 2017.

[165] Suman Saha, Gurkirt Singh, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. **Deep learning for detecting multiple space-time action tubes in videos**. *arXiv preprint arXiv:1608.01529*, 2016.

[166] Nathan Inkawhich, Matthew Inkawhich, Yiran Chen, and Hai Li. **Adversarial Attacks for Optical Flow-Based Action Recognition Classifiers**. *arXiv preprint arXiv:1811.11875*, 2018.

[167] Shuyang Sun, Zhanghui Kuang, Lu Sheng, Wanli Ouyang, and Wei Zhang. **Optical flow guided feature: a fast and robust motion representation for video action recognition**. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, **8**, 2018.

[168] Zhenyang Li, Kirill Gavrilyuk, Efstratios Gavves, Mihir Jain, and Cees GM Snoek. **VideoLSTM convolves, attends and flows for action recognition**. *Computer Vision and Image Understanding*, **166**:41–50, 2018.

[169] Hossein Rahmani, Ajmal Mian, and Mubarak Shah. **Learning a deep model for human action recognition from novel viewpoints**. *IEEE transactions on pattern analysis and machine intelligence*, **40**(3):667–681, 2018.

[170] Shugao Ma, Jianming Zhang, Stan Sclaroff, Nazli Ikizler-Cinbis, and Leonid Sigal. **Space-time tree ensemble for action recognition and localization**. *International Journal of Computer Vision*, **126**(2-4):314–332, 2018.

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

[172] Piotr Dollár, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. **Behavior recognition via sparse spatio-temporal features**. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72. IEEE, 2005.

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

[174] Hossein Rahmani, Arif Mahmood, Du Q Huynh, and Ajmal Mian. **HOPC: Histogram of oriented principal components of 3D pointclouds for action recognition**. In *European conference on computer vision*, pages 742–757. Springer, 2014.

[175] Hamed Pirsiavash and Deva Ramanan. **Parsing videos of actions with segmental grammars**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 612–619, 2014.

[176] Hilde Kuehne, Ali Arslan, and Thomas Serre. **The language of actions: Recovering the syntax and semantics of goal-directed human activities**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 780–787, 2014.

[177] Hilde Kuehne, Juergen Gall, and Thomas Serre. **An end-to-end generative framework for video segmentation and recognition**. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–8. IEEE, 2016.

[178] Gunnar A Sigurdsson, Santosh Kumar Divvala, Ali Farhadi, and Abhinav Gupta. **Asynchronous Temporal Fields for Action Recognition.** In *CVPR*, **5**, page 7, 2017.

[179] Effrosyni Mavroudi, Divya Bhaskara, Shahin Sefati, Haider Ali, and René Vidal. **End-to-End Fine-Grained Action Segmentation and Recognition Using Conditional Random Field Models and Discriminative Sparse Coding**. *arXiv preprint arXiv:1801.09571*, 2018.

[180] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu. **Pixel recurrent neural networks**. *arXiv preprint arXiv:1601.06759*, 2016.

[181] Atanas Mirchev and Seyed-Ahmad Ahmadi. **Classification of sparsely labeled spatio-temporal data through semi-supervised adversarial learning**. *arXiv preprint arXiv:1801.08712*, 2018.

[182] Jianting Fu, Liang Xiong, Xiaoying Song, Zhuo Yan, and Yi Xie. **Identification of finger movements from forearm surface EMG using an augmented probabilistic neural network**. In *System Integration (SII), 2017 IEEE/SICE International Symposium on*, pages 547–552. IEEE, 2017.

[183] Alexander Richard, Hilde Kuehne, and Juergen Gall. **Action Sets: Weakly Supervised Action Segmentation without Ordering Constraints**. *arXiv preprint arXiv:1706.00699*, 2017.

[184] Samuel Dixon. **Human Activity Workflow Parsing**. 2018.

[185] Liangliang Wang, Lianzheng Ge, Ruifeng Li, and Yajun Fang. **Three-stream CNNs for action recognition**. *Pattern Recognition Letters*, **92**:33–40, 2017.

[186] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. **Learning actionlet ensemble for 3D human action recognition**. *IEEE transactions on pattern analysis and machine intelligence*, **36**(5):914–927, 2014.

[187] Raviteja Vemulapalli and Rama Chellapa. **Rolling rotations for recognizing human actions from 3d skeletal data**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4471–4479, 2016.

[188] Novanto Yudistira and Takio Kurita. **Deep Packet Flow: Action Recognition via Multiresolution Deep Wavelet Packet of Local Dense Optical Flows**. *Journal of Signal Processing Systems*, pages 1–17, 2018.

[189] Mahesh Goyani and Narendra Patel. **Multi-Level Haar Wavelet based Facial Expression Recognition using Logistic Regression**. *Indian Journal of Science and Technology*, **10**(9), 2017.

[190] Yemin Shi, Yonghong Tian, Yaowei Wang, and Tiejun Huang. **Sequential deep trajectory descriptor for action recognition with three-stream CNN**. *IEEE Transactions on Multimedia*, **19**(7):1510–1520, 2017.

[191] Gül Varol, Ivan Laptev, and Cordelia Schmid. **Long-term temporal convolutions for action recognition**. *IEEE transactions on pattern analysis and machine intelligence*, **40**(6):1510–1517, 2018.

[192] Amr Sharaf, Marwan Torki, Mohamed E Hussein, and Motaz El-Saban. **Real-time multi-scale action detection from 3D skeleton data**. In *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*, pages 998–1005. IEEE, 2015.

[193] Rizwan Chaudhry, Ferda Ofli, Gregorij Kurillo, Ruzena Bajcsy, and Rene Vidal. **Bio-inspired dynamic 3d discriminative skeletal features for human action recognition**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 471–478, 2013.

[194] Eshed Ohn-Bar and Mohan Trivedi. **Joint angles similarities and HOG2 for action recognition**. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 465–470, 2013.

[195] Zhe Lin, Zhuolin Jiang, and Larry S Davis. **Recognizing actions by shape-motion prototype trees**. In *2009 IEEE 12th international conference on computer vision*, pages 444–451. IEEE, 2009.

[196] Ling Shao, Ling Ji, Yan Liu, and Jianguo Zhang. **Human action segmentation and recognition via motion and shape analysis**. *Pattern Recognition Letters*, **33**(4):438–445, 2012.

[197] Mohiuddin Ahmad and Seong-Whan Lee. **Human action recognition using shape and CLG-motion flow from multi-view image sequences**. *Pattern Recognition*, **41**(7):2237–2252, 2008.

[198] An-An Liu, Ning Xu, Wei-Zhi Nie, Yu-Ting Su, and Yong-Dong Zhang. **Multi-Domain and Multi-Task Learning for Human Action Recognition**. *IEEE Transactions on Image Processing*, **28**(2):853–867, 2019.

[199] Amin Ullah, Khan Muhammad, Ijaz Ul Haq, and Sung Wook Baik. **Action recognition using optimized deep autoencoder and CNN for surveillance data streams of non-stationary environments**. *Future Generation Computer Systems*, **96**:386–397, 2019.

[200] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. **Illumination-aware faster R-CNN for robust multispectral pedestrian detection**. *Pattern Recognition*, **85**:161–171, 2019.

[201] Yanpeng Cao, Dayan Guan, Weilin Huang, Jiangxin Yang, Yanlong Cao, and Yu Qiao. **Pedestrian detection with unsupervised multispectral feature learning using deep neural networks**. *Information Fusion*, **46**:206–217, 2019.

[202] Haidar A AlMubarak, Joe Stanley, Peng Guo, Rodney Long, Sameer Antani, George Thoma, Rosemary Zuna, Shelliane Frazier, and William Stoecker. **A Hybrid Deep Learning and Handcrafted Feature Approach for Cervical Cancer Digital Histology Image Classification**. *International Journal of Healthcare Information Systems and Informatics (IJHISI)*, **14**(2):66–87, 2019.

[203] Jinxing Li, Bob Zhang, Guangming Lu, and David Zhang. **Generative multi-view and multi-feature learning for classification**. *Information Fusion*, **45**:215–226, 2019.

[204] Wenqing Chu, Hongyang Xue, Chengwei Yao, and Deng Cai. **Sparse Coding Guided Spatiotemporal Feature Learning for Abnormal Event Detection in Large Videos**. *IEEE Transactions on Multimedia*, **21**(1):246–255, 2019.

[205] Li Liu, Ling Shao, Xuelong Li, and Ke Lu. **Learning spatio-temporal representations for action recognition: A genetic programming approach**. *IEEE transactions on cybernetics*, **46**(1):158–170, 2016.

REFERENCES

[206] Wei-Ta Chu and Hao-An Chu. **A Genetic Programming Approach to Integrate Multilayer CNN Features for Image Classification**. In *International Conference on Multimedia Modeling*, pages 640–651. Springer, 2019.

[207] Tingting Yao, Zhiyong Wang, Zhao Xie, Jun Gao, and David Dagan Feng. **Learning universal multiview dictionary for human action recognition**. *Pattern Recognition*, **64**:236–244, 2017.

[208] Ioannis Mademlis, Anastasios Tefas, and Ioannis Pitas. **Greedy salient dictionary learning for activity video summarization**. In *International Conference on Multimedia Modeling*, pages 578–589. Springer, 2019.

[209] Yongqiang Li, S Mohammad Mavadati, Mohammad H Mahoor, Yongping Zhao, and Qiang Ji. **Measuring the intensity of spontaneous facial action units with dynamic Bayesian network**. *Pattern Recognition*, **48**(11):3417–3427, 2015.

[210] Li Liu, Shu Wang, Bin Hu, Qingyu Qiong, Junhao Wen, and David S Rosenblum. **Learning structures of interval-based Bayesian networks in probabilistic generative model for human complex activity recognition**. *Pattern Recognition*, **81**:545–561, 2018.

[211] Fan Zhu and Ling Shao. **Weakly-supervised cross-domain dictionary learning for visual recognition**. *International Journal of Computer Vision*, **109**(1-2):42–59, 2014.

[212] James Gleick and Freeman J Dyson. **Genius: The Life and Science of Richard Feynman**. *Physics Today*, **45**:87, 1992.

[213] LI Xing and XIAO Qin-kun. **Human Action Recognition Using Auto-encode and PNN Neural Network**. *Software Guide*, (1):4, 2018.

[214] Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. **Unsupervised learning of video representations using lstms**. In *International conference on machine learning*, pages 843–852, 2015.

[215] Carl Doersch. **Tutorial on variational autoencoders**. *arXiv preprint arXiv:1606.05908*, 2016.

[216] Mahdyar Ravanbakhsh, Moin Nabi, Enver Sangineto, Lucio Marcenaro, Carlo Regazzoni, and Nicu Sebe. **Abnormal event detection in videos using generative adversarial nets**. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1577–1581. IEEE, 2017.

[217] Mohammad Sabokrou, Mohammad Khalooei, Mahmood Fathy, and Ehsan Adeli. **Adversarially Learned One-Class Classifier for Novelty Detection**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3379–3388, 2018.

[218] Michael Mathieu, Camille Couprie, and Yann LeCun. **Deep multi-scale video prediction beyond mean square error**. *arXiv preprint arXiv:1511.05440*, 2015.

[219] Maria Cornacchia, Koray Ozcan, Yu Zheng, and Senem Velipasalar. **A survey on activity detection and classification using wearable sensors**. *IEEE Sensors Journal*, **17**(2):386–403, 2017.

[220] Rajesh Kumar Tripathi, Anand Singh Jalal, and Subhash Chand Agrawal. **Suspicious human activity recognition: a review**. *Artificial Intelligence Review*, pages 1–57, 2017.

[221] Antonis A Argyros and Manolis IA Lourakis. **Binocular hand tracking and reconstruction based on 2D shape matching**. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, **1**, pages 207–210. IEEE, 2006.

[222] Michalis Vrigkas, Christophoros Nikou, and Ioannis A Kakadiadis. **Classifying behavioral attributes using conditional random fields**. In *Hellenic Conference on Artificial Intelligence*, pages 95–104. Springer, 2014.

[223] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. **Learning multimodal latent attributes**. *IEEE transactions on pattern analysis and machine intelligence*, **36**(2):303–316, 2014.

[224] Hyun Soo Park and Jianbo Shi. **Social saliency prediction**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4777–4785, 2015.

[225] Yuwei Wu, Yuanquan Wang, and Yunde Jia. **Adaptive diffusion flow active contours for image segmentation**. *Computer Vision and Image Understanding*, **117**(10):1421–1435, 2013.

[226] Soulan Liu, Chen Chen, and Nasser Kehtarnavaz. **A computationally efficient denoising and hole-filling method for depth image enhancement**. In *Real-Time Image and Video Processing 2016*, **9897**, page 98970V. International Society for Optics and Photonics, 2016.

[227] Weiyao Lin, Yang Mi, Weiyue Wang, Jianxin Wu, Jingdong Wang, and Tao Mei. **A diffusion and clustering-based approach for finding coherent motions and understanding crowd scenes**. *IEEE Transactions on Image Processing*, **25**(4):1674–1687, 2016.

[228] Jianbo Shi and Carlo Tomasi. **Good Features to Track**. In *1994 Proceedings of IEEE conference on computer vision and pattern recognition*, pages 593–600. IEEE, 1994.

[229] Bruce D. Lucas and Takeo Kanade. **An Iterative Image Registration Technique with an Application to Stereo Vision**. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, pages 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc.

[230] Ying Wu and Thomas S Huang. **View-independent recognition of hand postures**. In *cvpr*, page 2088. IEEE, 2000.

[231] Ying Wu, John Y Lin, and Thomas S Huang. **Capturing natural hand articulation**. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, **2**, pages 426–432. IEEE, 2001.

[232] Anjum Ali and JK Aggarwal. **Segmentation and recognition of continuous human activity**. In *Detection and recognition of events in video, 2001. Proceedings. IEEE Workshop on*, pages 28–35. IEEE, 2001.

[233] Yi Li. **Hand gesture recognition using Kinect**. In *Software Engineering and Service Science (ICSESS), 2012 IEEE 3rd International Conference on*, pages 196–199. IEEE, 2012.

[234] Saad Ali and Mubarak Shah. **Human action recognition in videos using kinematic features and multiple instance learning**. *IEEE transactions on pattern analysis and machine intelligence*, **32**(2):288–303, 2010.

[235] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. **Show and tell: A neural image caption generator**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[236] Pushpajit Khaire, Praveen Kumar, and Javed Imran. **Combining CNN streams of RGB-D and skeletal data for human activity recognition**. *Pattern Recognition Letters*, 2018.

[237] Fernando Moya Rueda and Gernot A Fink. **Learning Attribute Representation for Human Activity Recognition**. *arXiv preprint arXiv:1802.00761*, 2018.

[238] Daniele Ravi, Charence Wong, Benny Lo, and Guang-Zhong Yang. **A deep learning approach to on-node sensor data analytics for mobile or wearable devices**. *IEEE journal of biomedical and health informatics*, **21**(1):56–64, 2017.

[239] Inês P Machado, A Luísa Gomes, Hugo Gamboa, Vítor Paixão, and Rui M Costa. **Human activity data discovery from triaxial accelerometer sensor: Non-supervised learning sensitivity to feature extraction parametrization**. *Information Processing & Management*, **51**(2):204–214, 2015.

[240] Gheorghe Sebestyen, Ionut Stoica, and Anca Hangan. **Human activity recognition and monitoring for elderly people**. In *Intelligent Computer Communication and Processing (ICCP), 2016 IEEE 12th International Conference on*, pages 341–347. IEEE, 2016.

[241] Yu Kong, Yunde Jia, and Yun Fu. **Interactive phrases: Semantic descriptionsfor human interaction recognition**. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (9):1775–1788, 2014.

[242] Qiuxia Wu, Zhiyong Wang, Feiqi Deng, Zheru Chi, and David Dagan Feng. **Realistic human action recognition with multimodal feature selection and fusion**. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, **43**(4):875–885, 2013.

[243] Maheshkumar H Kolekar and Deba Prasad Dash. **Hidden markov model based human activity recognition using shape and optical flow based features**. In *2016 IEEE Region 10 Conference (TENCON)*, pages 393–397. IEEE, 2016.

[244] Katarzyna Gościewska and Dariusz Frejlichowski. **Silhouette-Based Action Recognition Using Simple Shape Descriptors**. In *International Conference on Computer Vision and Graphics*, pages 413–424. Springer, 2018.

[245] Amin Ullah, Jamil Ahmad, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik. **Action recognition in video sequences using deep bi-directional LSTM with CNN features**. *IEEE Access*, **6**:1155–1166, 2017.

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

[247] Zhiwei Deng, Arash Vahdat, Hexiang Hu, and Greg Mori. **Structure inference machines: Recurrent neural networks for analyzing relations in group activity recognition**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4772–4781, 2016.

[248] Chen Chen, Mengyuan Liu, Hong Liu, Baochang Zhang, Jungong Han, and Nasser Kehtarnavaz. **Multi-temporal depth motion maps-based local binary patterns for 3-D human action recognition**. *IEEE Access*, **5**:22590–22604, 2017.

[249] Hossein Rahmani and Ajmal Mian. **3D action recognition from novel viewpoints**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1506–1515, 2016.

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

[251] Pengfei Zhang, Cuiling Lan, Junliang Xing, Wenjun Zeng, Jianru Xue, and Nanning Zheng. **View adaptive recurrent neural networks for high performance human action recognition from skeleton data**. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2117–2126, 2017.

[252] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. **An end-to-end spatio-temporal attention model for human action recognition from skeleton data**. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[253] Mengyuan Liu, Hong Liu, and Chen Chen. **Enhanced skeleton visualization for view invariant human action recognition**. *Pattern Recognition*, **68**:346–362, 2017.

[254] Soumitra Samanta and Bhabatosh Chanda. **Space-time facet model for human activity classification**. *IEEE Transactions on Multimedia*, **16**(6):1525–1535, 2014.

[255] Gang Yu and Junsong Yuan. **Fast action proposals for human action detection and search**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1302–1311, 2015.

[256] Vignesh Ramanathan, Congcong Li, Jia Deng, Wei Han, Zhen Li, Kunlong Gu, Yang Song, Samy Bengio, Charles Rosenberg, and Li Fei-Fei. **Learning semantic relationships for better action retrieval in images**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1100–1109, 2015.

[257] Yong Du, Wei Wang, and Liang Wang. **Hierarchical recurrent neural network for skeleton based action recognition**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1110–1118, 2015.

[258] Chenxia Wu, Jiemi Zhang, Ozan Sener, Bart Selman, Silvio Savarese, and Ashutosh Saxena. **Watch-n-patch: unsupervised learning of actions and relations**. *IEEE transactions on pattern analysis and machine intelligence*, **40**(2):467–481, 2017.

[259] J. Liu, R. Tan, G. Han, N. Sun, and S. Kwong. **Privacy-preserving In-home Fall Detection Using Visual Shielding Sensing and Private Information-embedding**. *IEEE Transactions on Multimedia*, pages 1–1, 2020.

[260] Yu Liu, John SY Chan, and Jin H Yan. **Neuropsychological mechanisms of falls in older adults**. *Frontiers in aging neuroscience*, **6**:64, 2014.

[261] Anita Ramachandran and Anupama Karuppiah. **A Survey on Recent Advances in Wearable Fall Detection Systems**. *BioMed Research International*, **2020**, 2020.

[262] **World Health Organization, WHO Global Report on Falls Prevention in Older Age**. Technical report, 2007.

[263] Nadia Oukrich. *Daily Human Activity Recognition in Smart Home based on Feature Selection, Neural Network and Load Signature of Appliances*. PhD thesis, 2019.

[264] Peter F Edemekong, Deb Bomgaars, Sukesh Sukumaran, and Shoshana B Levy. **Activities of daily living**. *StatPearls [Internet]. Treasure Island FL. StatPearls Publishing*, 2020.

[265] Jung Keun Lee, Stephen N Robinovitch, and Edward J Park. **Inertial sensing-based pre-impact detection of falls involving near-fall scenarios**. *IEEE transactions on neural systems and rehabilitation engineering*, **23**(2):258–266, 2014.

[266] Soonjae Ahn, Jongman Kim, Bummo Koo, and Youngho Kim. **Evaluation of inertial sensor-based pre-impact fall detection algorithms using public dataset**. *Sensors*, **19**(4):774, 2019.

REFERENCES

[267] Angelo Maria Sabatini, Gabriele Ligorio, Andrea Mannini, Vincenzo Genovese, and Laura Pinna. **Prior-to-and post-impact fall detection using inertial and barometric altimeter measurements**. *IEEE transactions on neural systems and rehabilitation engineering*, **24**(7):774–783, 2015.

[268] GE Wu and Shuwan Xue. **Portable preimpact fall detector with inertial sensors**. *IEEE Transactions on neural systems and Rehabilitation Engineering*, **16**(2):178–183, 2008.

[269] Nashwa El-Bendary, Qing Tan, Frédérique C Pivot, and Anthony Lam. **FALL DETECTION AND PREVENTION FOR THE ELDERLY: A REVIEW OF TRENDS AND CHALLENGES.** *International Journal on Smart Sensing & Intelligent Systems*, **6**(3), 2013.

[270] J Klenk, Clemens Becker, F Lieken, S Nicolai, W Maetzler, W Alt, W Zijlstra, JM Hausdorff, RC Van Lummel, L Chiari, et al. **Comparison of acceleration signals of simulated and real-world backward falls**. *Medical engineering & physics*, **33**(3):368–373, 2011.

[271] Shehroz S Khan and Jesse Hoey. **Review of fall detection techniques: A data availability perspective**. *Medical engineering & physics*, **39**:12–22, 2017.

[272] Xin Ma, Haibo Wang, Bingxia Xue, Mingang Zhou, Bing Ji, and Yibin Li. **Depth-based human fall detection via shape features and improved extreme learning machine**. *IEEE journal of biomedical and health informatics*, **18**(6):1915–1922, 2014.

[273] Georgios Goudelis, Georgios Tsatiris, Kostas Karpouzis, and Stefanos Kollias. **Fall detection using history triple features**. In *Proceedings of the 8th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, pages 1–7, 2015.

[274] Glen Debard, Marc Mertens, Mieke Deschodt, Ellen Vlaeyen, Els Devriendt, Eddy Dejaeger, Koen Milisen, Jos Tournoy, Tom Croonenborghs, Toon Goedemé, et al. **Camera-based fall detection using real-world versus simulated data: How far are we from the solution?** *Journal of Ambient Intelligence and Smart Environments*, **8**(2):149–168, 2016.

[275] Weiguo Feng, Rui Liu, and Ming Zhu. **Fall detection for elderly person care in a vision-based home surveillance environment using a monocular camera**. *signal, image and video processing*, **8**(6):1129–1138, 2014.

[276] Ahmet Iscen, Anil Armagan, and Pinar Duygulu. **What is usual in unusual videos? Trajectory snippet histograms for discovering unusualness**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 794–799, 2014.

[277] Viet Anh Nguyen, Thanh Ha Le, and Thuy Thi Nguyen. **Single camera based fall detection using motion and human shape features**. In *Proceedings of the Seventh Symposium on Information and Communication Technology*, pages 339–344, 2016.

[278] Yixiao Yun and Irene Yu-Hua Gu. **Human fall detection in videos via boosting and fusing statistical features of appearance, shape and motion dynamics on Riemannian manifolds with applications to assisted living**. *Computer Vision and Image Understanding*, **148**:111–122, 2016.

[279] Smriti Bhandari, Navnee Babar, Pranav Gupta, Nidhi Shah, and Shreyas Pujari. **A novel approach for fall detection in home environment**. In *2017 IEEE 6th Global Conference on Consumer Electronics (GCCE)*, pages 1–5. IEEE, 2017.

[280] Pooja Shukla and Arti Tiwari. **Vision based approach to human fall detection**. *International Journal of Engineering Research and General Science*, **3**(6):944–949, 2015.

[281] Caroline Rougier, Alain St-Arnaud, Jacqueline Rousseau, and Jean Meunier. **Video surveillance for fall detection**. *Video Surveillance*, **21**(5):611–622, 2011.

[282] Kamal Sehairi, Fatima Chouireb, and Jean Meunier. **Elderly fall detection system based on multiple shape features and motion analysis**. In *2018 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pages 1–8. IEEE, 2018.

[283] Kaibo Fan, Ping Wang, Yan Hu, and Bingjie Dou. **Fall detection via human posture representation and support vector machine**. *International journal of distributed sensor networks*, **13**(5):1550147717707418, 2017.

[284] Georgios Mastorakis and Dimitrios Makris. **Fall detection system using Kinect's infrared sensor**. *Journal of Real-Time Image Processing*, **9**(4):635–646, 2014.

[285] Yanfei Peng, Jianjun Peng, Jiping Li, Pitao Yan, and Bing Hu. **Design and development of the fall detection system based on point cloud**. *Procedia computer science*, **147**:271–275, 2019.

[286] L Ciabattoni, G Foresi, A Monteriù, D Proietti Pagnotta, and L Tomaiuolo. **Fall detection system by using ambient intelligence and mobile robots**. In *2018 Zooming Innovation in Consumer Technologies Conference (ZINC)*, pages 130–131. IEEE, 2018.

[287] Adrian Nunez-Marcos, Gorka Azkune, and Ignacio Arganda-Carreras. **Vision-based fall detection with convolutional neural networks**. *Wireless communications and mobile computing*, **2017**, 2017.

[288] Lesya Anishchenko. **Machine learning in video surveillance for fall detection**. In *2018 Ural Symposium on Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT)*, pages 99–102. IEEE, 2018.

[289] Qing Han, Haoyu Zhao, Weidong Min, Hao Cui, Xiang Zhou, Ke Zuo, and Ruikang Liu. **A Two-Stream Approach to Fall Detection With MobileVGG**. *IEEE Access*, **8**:17556–17566, 2020.

[290] Na Lu, Yidan Wu, Li Feng, and Jinbo Song. **Deep learning for fall detection: Three-dimensional CNN combined with LSTM on video kinematic data**. *IEEE journal of biomedical and health informatics*, **23**(1):314–323, 2018.

[291] Kripesh Adhikari, Hamid Bouchachia, and Hammadi Nait-Charif. **Long short-term memory networks based fall detection using unified pose estimation**. In *Twelfth International Conference on Machine Vision (ICMV 2019)*, **11433**, page 114330H. International Society for Optics and Photonics, 2020.

[292] Lourdes Martínez-Villaseñor, Hiram Ponce, Jorge Brieva, Ernesto Moya-Albor, José Núñez-Martínez, and Carlos Peñafort-Asturiano. **UP-fall detection dataset: A multimodal approach**. *Sensors*, **19**(9):1988, 2019.

[293] Korbinian Frank, Maria Josefa Vera Nadales, Patrick Robertson, and Tom Pfeifer. **Bayesian recognition of motion related activities with inertial sensors**. In *Proceedings of the 12th ACM international conference adjunct papers on Ubiquitous computing-Adjunct*, pages 445–446, 2010.

[294] G Vavoulas, M Pediaditis, C Chatzaki, EG Spanakis, and M Tsiknakis. **Artificial intelligence: concepts, methodologies, tools, and applications**, 2017.

[295] Carlos Medrano, Raul Igual, Inmaculada Plaza, and Manuel Castro. **Detecting falls as novelties in acceleration patterns acquired with smartphones**. *PloS one*, **9**(4):e94811, 2014.

[296] Thomas Vilarinho, Babak Farshchian, Daniel Gloppestad Bajer, Ole Halvor Dahl, Iver Egge, Sondre Steinsland Hegdal, Andreas Lønes, Johan N Slettevold, and Sam Mathias Weggersen. **A combined smartphone and smartwatch fall detection system**. In *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing*, pages 1443–1448. IEEE, 2015.

[297] Eduardo Casilari, Jose A Santoyo-Ramón, and Jose M Cano-García. **Umafall: A multisensor dataset for the research on automatic fall detection**. *Procedia Computer Science*, **110**:32–39, 2017.

[298] Angela Sucerquia, José David López, and Jesús Francisco Vargas-Bonilla. **SisFall: A fall and movement dataset**. *Sensors*, **17**(1):198, 2017.

[299] Fairouz Merrouche and Nadia Baha. **Fall detection based on shape deformation**. *Multimedia Tools and Applications*, **79**(41):30489–30508, 2020.

[300] Imen Charfi, Johel Miteran, Julien Dubois, Mohamed Atri, and Rached Tourki. **Optimized spatio-temporal descriptors for real-time fall detection: comparison of support vector machine and Adaboost-based classification**. *Journal of Electronic Imaging*, **22**(4):041106, 2013.

[301] Zhong Zhang, Christopher Conly, and Vassilis Athitsos. **Evaluating depth-based computer vision methods for fall detection under occlusions**. In *International Symposium on Visual Computing*, pages 196–207. Springer, 2014.

[302] Bogdan Kwolek and Michal Kepski. **Human fall detection on embedded platform using depth maps and wireless accelerometer**. *Computer methods and programs in biomedicine*, **117**(3):489–501, 2014.

[303] Edouard Auvinet, Caroline Rougier, Jean Meunier, Alain St-Arnaud, and Jacqueline Rousseau. **Multiple cameras fall dataset**. *DIRO-Université de Montréal, Tech. Rep*, **1350**, 2010.

[304] Annalisa Franco, Antonio Magnani, and Dario Maio. **A multimodal approach for human activity recognition based on skeleton and RGB data**. *Pattern Recognition Letters*, 2020.

[305] Mohammad Haghighat, Mohamed Abdel-Mottaleb, and Wadee Alhalabi. **Fully automatic face normalization and single sample face recognition in unconstrained environments**. *Expert Systems with Applications*, **47**:23–34, 2016.

[306] Girija Chetty and Matthew White. **Body sensor networks for human activity recognition**. In *2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 660–665. IEEE, 2016.

[307] Basura Fernando, Efstratios Gavves, José Oramas, Amir Ghodrati, and Tinne Tuytelaars. **Rank pooling for action recognition**. *IEEE transactions on pattern analysis and machine intelligence*, **39**(4):773–787, 2016.

[308] HA Sial, MH Yousaf, and F Hussain. **Spatio-Temporal RGBD Cuboids Feature for Human Activity Recognition**. *The Nucleus*, **55**(3):139–149, 2018.

[309] Rishi Chopra et al. **ACTIVITY RECOGNITION BASED ON 3D CNN-LSTM-ASSISTED APPROACH**. *Journal of the Gujarat Research Society*, **21**(6):454–466, 2019.

[310] Marco Leo, Nicola Mosca, Paolo Spagnolo, Pier Luigi Mazzeo, Tiziana D'Orazio, and Arcangelo Distante. **Real-time multiview analysis of soccer matches for understanding interactions between ball and players**. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 525–534, 2008.

[311] Basura Fernando, Peter Anderson, Marcus Hutter, and Stephen Gould. **Discriminative hierarchical rank pooling for activity recognition**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1924–1932, 2016.

[312] Snehasis Mukherjee, Leburu Anvitha, and T Mohana Lahari. **Human Activity Recognition in RGB-D Videos by Dynamic Images**. *arXiv preprint arXiv:1807.02947*, 2018.

[313] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. **UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor**. In *2015 IEEE International conference on image processing (ICIP)*, pages 168–172. IEEE, 2015.

[314] Mohammad Farhad Bulbul, Yunsheng Jiang, and Jinwen Ma. **DMMs-based multiple features fusion for human action recognition**. *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, **6**(4):23–39, 2015.

[315] Pushpajit Khaire, Javed Imran, and Praveen Kumar. **Human Activity Recognition by Fusion of RGB, Depth, and Skeletal Data**. In *Proceedings of 2nd International Conference on Computer Vision & Image Processing*, pages 409–421. Springer, 2018.

[316] Javed Imran and Balasubramanian Raman. **Evaluating fusion of RGB-D and inertial sensors for multimodal human action recognition**. *Journal of Ambient Intelligence and Humanized Computing*, **11**(1):189–208, 2020.

[317] Pichao Wang, Zhaoyang Li, Yonghong Hou, and Wanqing Li. **Action recognition based on joint trajectory maps using convolutional neural networks**. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 102–106, 2016.

[318] Rogmans W Tuner S, Kisser R. **Falls among older adults in the EU-28: key facts from the available statistics. Másinformación: www.falls kip.com**. Technical report, 2015.

[319] CS Florence, G Bergen, A Atherly, E Burns, J Stevens, and C Drake. **Medical costs of fatal and nonfatal falls in older adults**. *Journal of the American Geriatrics Society*, **66**(4):693–698, 2018.

[320] C Wang, W Lu, SJ Redmond, MC Stevens, SR Lord, and NH Lovell. **A low-power fall detector balancing sensitivity and false alarm rate**. *IEEE Journal of Biomedical Health Informatics*, **22**(6):1929–1937, 2018.

[321] Engin Mendi, Hélio B Clemente, and Coskun Bayrak. **Sports video summarization based on motion analysis**. *Computers & Electrical Engineering*, **39**(3):790–796, 2013.

[322] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. **Convolutional pose machines**. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 4724–4732, 2016.

[323] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. **2D Human Pose Estimation: New Benchmark and State of the Art Analysis**. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[324] Philippe Remy. **Temporal Convolutional Networks for Keras**. `https://github.com/philipperemy/keras-tcn`, 2020.

[325] Nabil Zerrouki and Amrane Houacine. **Combined curvelets and hidden Markov models for human fall detection**. *Multimedia Tools and Applications*, **77**(5):6405–6424, 2018.

REFERENCES

[326] Bo-Hua Wang, Jie Yu, Kuo Wang, Xuan-Yu Bao, and Ke-Ming Mao. **Fall detection based on dual-channel feature integration**. *IEEE Access*, **8**:103443–103453, 2020.

[327] Vincenzo Dentamaro, Donato Impedovo, and Giuseppe Pirlo. **Fall Detection by Human Pose Estimation and Kinematic Theory**. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 2328–2335. IEEE, 2021.

[328] Shabnam Ezatzadeh and Mohammad Reza Keyvanpour. **Fall detection for elderly in assisted environments: Video surveillance systems and challenges**. In *2017 9th International Conference on Information and Knowledge Technology (IKT)*, pages 93–98. IEEE, 2017.

[329] IPES Putra, James Brusey, Elena Gaura, and Rein Vesilo. **An event-triggered machine learning approach for accelerometer-based fall detection**. *Sensors*, **18**(1):20, 2018.

Appendix A

Table A.1: Analysis of some state-of-the-art comprehensive surveys on HAR

Paper	Year	Contribution	Application field	Activities type	Body parts	Image vs video	Limitations
(10)	2015	Gesture taxonomies, hand gesture recognition applications and approaches.	Different fields	Gestures	Hands	Image and video	Not interested to different activities and focuses only on gestures.
(22)	2015	An approach-based taxonomy of the state-of-the-art research and advances in HAR.	Different fields	Different types	Different body parts	Image and video	Follows a specific taxonomy and doesn't cover a wide range of deep-learning based aproaches.
(23)	2015	Comprehensive survey on kinect-based motion recognition techniques and the underlying datasets.	Applications of the Kinect technology	Different types	Different body parts	Image and video	Devoted only to motion recognition using data captured by Kinect.
(24)	2015	Classification of HAR approaches regarding the source of input data into unimodal or multimodal methods and analysis of some publicly available human activity datasets.	Different fields	Different types	Different body parts	Image and video	Presents an approach-based taxonomy according to the source of input data but doesn't cover many general aspects of HAR.
(25)	2015	Semantic-based human recognition methods and a brief representation of their application fields.	Different fields	Actions and interac-tions	Different body parts	Image and video	Focuses on semantic-based HAR methods and doesn't include many general aspects of HAR.

Continued on next page

Table A.1 – continued from previous page

Paper	Year	Contribution	Application field	Activities type	Body parts	Image vs video	Limitations
(144)	2015	Video-based HAR using deep learning and classification of datasets according to different complexity levels.	Different fields	Different types	Different body parts	Image and video	Interested to deep learning based HAR and doesn't cover hand-crafted approaches.
(2)	2016	Analysis of different HAR methods and comparison between different action identification methods.	Different fields	Different types	Different body parts	Video	Doesn't cover many HAR approaches and general aspects.
(26)	2016	Comprehensive survey on the recent techniques of HAR.	Different fields	Actions	Different body parts	Video	Doesn't cover many aspects and benchmark databases of HAR.
(27)	2016	Classification of various action recognition and detection algorithms according to the extraction and encoding of features, as well as the classification processes.	Different fields	Different types	Different body parts	Image and video	Covers many aspects of HAR but many research works have emerged from 2016 till now.
(28)	2016	Overview of methodologies, challenges and issues of HAR systems.	Different fields	Actions and interac-tions	Different body parts	Image and video	Doesn't provide an indepth study. It doesn't cover many techniques and aspects of HAR.
(112)	2016	3D skeleton-based HAR approaches.	Different fields	Actions	Different body parts	Video	Focuses mostly on 3D skeleton-based HAR and omits a wide range of other approaches.
(114)	2016	Analysis of popular techniques used for object segmentation and recognition.	Different fields	Actions	Different body parts	Video	Devoted to object segmentation and detection in general and is not specific to HAR.
(46)	2016	Overview of HAR techniques in videos.	Surveillance, entertain-ment and healthcare.	Different types	Different body parts	Video	Doesn't cover many HAR approaches and is limited to some specific application fields.

Continued on next page

Table A.1 – continued from previous page

Paper	Year	Contribution	Application field	Activities type	Body parts	Image vs video	Limitations
(115)	2016	Comprehensive survey on the recent development and challenges of human detection.	Different fields	Actions	Different body parts	Image and video	Devoted to human activities detection and doesn't cover the whole process of HAR.
(116)	2016	Knowledge-based HAR methodologies.	Different fields	Actions	Different body parts	Video	Interested to methods incorporating a priori knowledge and context information on the activity and doesn't cover many other approaches.
(120)	2016	Comprehensive survey of the emerging progress on 3D hand gesture recognition approaches and systems.	Human computer interaction	Gestures	Hands	Video	The emphasis is on 3D hand gesture recognition approaches. It doesn't cover other activity types.
(143)	2016	Comprehensive analysis and comparison between learning-based and handcrafted action representations.	Different fields	Different types	Different body parts	Image and video	Presents a human action representation based taxonomy and omits many other aspects of HAR.
(152)	2016	Current state of publicly available HAR datasets.	Different fields	Different types	Different body parts	Image and video	The focus of this survey is the available datasets for HAR. It doesn't cover HAR approaches.
(156)	2016	Comprehensive review of the most commonly used action recognition related RGB-D video datasets.	Different fields	Different types	Different body parts	Video	Presents only RGB-D video datasets and doesn't discuss HAR approaches.
(60)	2016	Review of the state of the art of vision-based systems for the recognition of daily life activities.	Daily life activities	Different types	Different body parts	Video	The focus is made on techniques related to daily life activities and omits many other application domains.

Continued on next page

Table A.1 – continued from previous page

Paper	Year	Contribution	Application field	Activities type	Body parts	Image vs video	Limitations
(29)	2017	Categorization of video-based HAR techniques into handcrafted feature-based and deep learning-based approaches.	Different fields	Actions	Different body parts	Video	The focus is automatic HAR techniques in videos. Still images are omitted.
(113)	2017	Survey of existing space-time action representations based on 3D skeletal data.	Different fields	Actions	Different body parts	Video	Focuses on 3D human representation based on skeletal data, and omits other data representations.
(121)	2017	State of the art of multimodal gesture recognition.	Machine learning and computer vision	Gestures	Hands	Image and video	Devoted to gesture recognition using multimodal data. It doesn't cover other activity types.
(128)	2017	Complex event recognition techniques.	Event recognition	Event	Different body parts	Image and video	Interested to complex event techniques and doesn't cover HAR approaches related to other activity types.
(49)	2017	Comprehensive review of the notable steps taken towards recognizing human actions.	Different fields	Actions	Different body parts	Image and video	Classifies methods of human action only and doesn't cover other activity types and action detection methods.
(145)	2017	Survey on current deep learning methodologies for action and gesture recognition.	Different fields	Actions and gestures	Different body parts	Video	Deep-learning based taxonomy for action and gesture recognition with particular interest on temporal dimension of data. Spatial features are not covered.

Continued on next page

Table A.1 – continued from previous page

Paper	Year	Contribution	Application field	Activities type	Body parts	Image vs video	Limitations
(131)	2017	Discussion of research works focusing on identifying, tracking and understanding group activities, interactions, and abnormal activities detection in large crowds with a summary of underlying available datasets.	Crowd management, public space design, and visual surveillance	Interactions and group activities	Different body parts	Video	Specific to crowd analysis for video surveillance purposes and abnormality detection. It doesn't cover other application domais.
(146)	2018	RGB-D-based human motion recognition with deep learning focusing on three architectures of neural networks.	Different fields	Different types	Different body parts	Video	Dedicated to RGB-D-based human motion recognition using deep learning and doesn't cover many other approaches.
(153) (158)	2019	Presentation and comparison of different types of video datasets, challenges, and their related latest evaluation techniques.	Different fields	Different types	Different body parts	Video	Devoted only to datasets analysis and doesn't discuss many other aspects of HAR.
(110)	2019	Survey of HAR methods, including progress in both hand-designed and deep learning-based action feature representation methods.	Different fields	Actions and interactions	Different body parts	Image and video	Doesn't cover many HAR approaches and different activity types.
(122)	2019	Review of state-of-the-art techniques used in recent hand gesture and sign language recognition research.	Sign language recognition	Gestures	Hands	Image and video	Devoted only to hand gesture recognition techniques and doesn't include other activity types.

Continued on next page

Table A.1 – continued from previous page

Paper	Year	Contribution	Application field	Activities type	Body parts	Image vs video	Limitations
(129)	2019	Summary of techniques of one person/ a group of people and their social interactions as well as their interaction with robots.	Robotics	Interactions and group activities	Different body parts	Video	Reviews recent approaches for recognizing human activities within the general framework of interaction with robots. It doesn't cover many other application domains.
(130)	2019	A comprehensive review on abnormal crowd behaviour detection methods.	Intelligent surveillance video systems	Interactions and group activities	Different body parts	Video	Focuses on detection abnormal activities methods in a crowded scene scenario. It omits many other application domains.
(132)	2019	Deep learning based techniques for various crowd video analysis methods.	Surveillance video analysis	Interactions and group activities	Different body parts	Video	Focuses on intelligent surveillance video analysis techniques and omits many other HAR application domains.
(133)	2019	State-of-the-art techniques on crowd behavior analysis, motion patterns, tracking, activity analysis and modeling. Evaluation metrics and datasets are also discussed.	Video surveillance	Interactions and group activities	Different body parts	Video	Devoted to crowd behavior analysis for video surveillance. It omits many other HAR approaches and application domains.
(134)	2019	Some insights of the state-of-the-art research works in the fields of Intelligent Video Surveillance, Wireless Sensor Network-based HAR, and camera-based health monitoring.	Intelligent Video Surveillance and health monitoring.	Actions	Different body parts	Video	Devoted to specific applications of HAR and doesn't cover many HAR approaches.

Continued on next page

Table A.1 – continued from previous page

Paper	Year	Contribution	Application field	Activities type	Body parts	Image vs video	Limitations
(135)	2019	In-depth and comprehensive analysis of data fusion and multiple techniques for HAR.	Different fields	Actions	Different body parts	Image and video	Emphasis on data fusion and applications of HAR on mobile and wearable devices. It omits many HAR approaches and applications.
(136)	2019	Overview of existing abnormal human activity recognition approaches.	Smart home surveillance and public place security	Actions, interac-tions and group activities	Different body parts	Image and video	Focuses on abnormal human activity recognition and fall detection. It omits many other applications domains and approaches of HAR.
(147)	2019	Comprehensive review of the CNN-based action recognition methods.	Different fields	Actions	Different body parts	Image and video	Presents CNN-based HAR approaches and doesn't include any of the handcrafted methods.

Declaration

I Beddiar Djamila Romaissa, herewith declare that I have produced this thesis titled "Vision-based Human Activity Recognition in Supervised or Assisted Environment" and the work presented in it is my own. I confirm that:

- My thesis is produced without the prohibited assistance of third parties and without making use of aids other than those specified.
- Ideas taken over directly or indirectly from other sources have been identified as such.
- I have acknowledged all main sources of help.
- The thesis work was conducted under the supervision of Professor Nini Brahim.
- This thesis has not previously been presented in identical or similar form to any other examination board.