Bi-lateral Interaction Between Humanoid Robots And Human

Zahra Ramezanpanah

université
PARIS-SACLAY

Bi-lateral Interaction Between Humanoid Robots And Human

Thèse de doctorat de l'Université Paris-Saclay

École doctorale n° 580 Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat: Robotique
Unité de recherche: IBISC EA 4526
Référent: : Malik Mallem

Thèse présentée et soutenue à Evry le 15 December 2020, par

Zahra RAMEZANPANAH

Composition du jury:

Pierre ZWEIGENBAUM DR CNRS, Université Paris-Saclay	Président
Catherine ACHARD Professeure, Sorbonne Université	Rapportrice
Titus ZAHARIA Professeur, Télécom SudParis/IPP	Rapporteur
Catherine PELACHAUD DR CNRS, Sorbonne Université	Examinatrice
Malik MALLEM Professeur, UEVE/Université Paris-Saclay	Directeur
Frédéric DAVESNE Ingénieur de recherche, UEVE/Université Paris-Saclay	Invité

Thèse de doctorat

NNT: 2020UPASG039

Acknowledgements

First of all, I would like to thank my thesis supervisor, Mr. Malik Mallem, for the confidence he placed in me by accepting to supervise this doctoral work, for his benevolence, for his numerous advices and for his availability despite his many tasks. Finally, I was extremely sensitive to his human qualities of listening and understanding throughout this doctoral work. I would also like to express my warmest thanks to my thesis co supervisor, Mr. Férédric Devesne, who supervised me throughout the years of my thesis. I would also like to tell him how much I appreciated his great availability and his unwavering respect for the tight deadlines for proofreading the documents I sent him. I would like to thank Mrs Catherine Achard and Mr Titus Zaharia for having accepted to be the reviewers of my thesis and Mrs Catherine Pelachaud and Mr Pierre Zweigenbaum and Frédéric Davesne for having examined this manuscript. I appreciate the interest you have shown in this work and thank you for devoting some of your precious time to me. A thesis is also a laboratory where you spend many hours and where it is important to feel good. So, I would like to thank all the members of the IBISC laboratory for their generosity and good humor. I would like to thank all the members of the IBISC laboratory for their generosity and good humor. From a more personal point of view, I thank my parents and my sister for always supporting me in my studies and in my choices. Despite the distance, they were always listening. I also thank my husband, Shahryar, for his patience, the encouragement and the attention he gave me during this thesis. I could hardly have finished my thesis in good conditions without his love and support. Finally, I would like to thank once again my dear parents, who probably could not be with me on graduation day because of this pandemic, and tell them that without their support, I could not have achieved any of my life goals.

Abstract

In this thesis, we address the issue of recognizing human body language in order to establish a bi-lateral interaction human-robot and robot-robot. New contributions have been made to this research. Our approach is founded on the identification of human gestures based on a motion analysis method that accurately describes motions. This thesis is divided into two parts: gesture recognition and emotion recognition based on the body gestures. In these two parts, we utilize two methods : classical Machine Learning and Deep Learning.

In the Gesture Recognition section, we first define a local descriptor based on the Laban Movement Analysis (LMA) to describe the movements. LMA is a method that uses four components to describe a movement: Body, Space, Shape and Effort. Since the only goal in this part is gesture recognition, only the first three factors are utilized. The Dynamic Time Warping (DTW) algorithm is implemented to find the similarities of the curves obtained from the descriptor vectors obtained by the LMA method. Finally, the Support Vector Machine, SVM, algorithm is utilized to train and classify the data. Thanks to normalization process, our system is invariant to the initial positions and orientations of people. By the use of Spline functions, the data are sampled in order to reduce the size of our descriptor and also to adapt the data to the classification methods. Several experiments are performed using public data sets.

In the second part of first section, we construct a new descriptor based on the geometric coordinates of different parts of the body in order to characterize a movement. To do this, in addition to the distances between hip center and other joints of the body and the angular changes, we define the triangles formed by the different parts of the body and calculated their area. We also calculate the area of the convex hell encompassing all the joints of the body. At the end we add the velocity of different joints in the proposed descriptor. We used a multi-layer long short-term memory (LSTM) network to evaluate this descriptor. The proposed algorithm is implemented on five public data sets, MRSAction 3D, Florence 3D, UTKinent, SYSU 3D HOINTU and RGB+D 120 data sets, and the results are compared with those available in the literature.

In the second section of this thesis, we first present a higher level algorithm to identify the inner feelings of human beings by observing their body movements. In order to define a robust descriptor, two methods are carried out: the first method is the LMA with the "Effort" factor, which describes a movement and the state in which it was performed. The second one is based on a set of spatio-temporal features. In the continuation of this section, a pipeline of expressive motions recognition is proposed in order to classify the emotions of people through their gestures by the use of machine learning methods (Random Decision Forest, Feed forward Neural Network). A comparative study is made between these two methods in order to choose the best one. The approach is validated with public data sets and our own data set of expressive gestures called Xsens Expressive

Motion (XEM).

In a second part of this section, we carry out of a statistical study based on human perception in order to evaluate the recognition system as well as the proposed motion descriptor. This allows us to estimate the capacity of our system to be able to classify and analyze human emotions. In this part two tasks are carried out with the two classifiers (the RDF for learning and the human approach for validation).

Keywords: Descriptors based on LMA, Gesture Recognition, Emotion based on body gestures, Human-robot interaction.

Contents

List of Figures

List of Tables

Chapter 1

General Introduction

Contents

1.1 General context

Nowadays, machines and, in particular, computers and robots have become an important part of our environment. They support our communications, perform difficult tasks for us, and ultimately they are meant to be our assistants. Therefore interactions with these machines appear to be a key problem so it must be treated carefully. Indeed, a machine is accepted and useful if (1) it is easy to program and control, (2) it achieves the expected results.These two demands are completely related to the naturalness, intuition and ease of communication between these devices and humans. Inspired by human-human interaction, body language, particularly gestures, is still omnipresent and essential. We use gestures without being aware of them, either to manipulate objects, or to communicate, transmit messages, express emotions, etc. In this context, the objective of our thesis is to develop a natural interaction between humans and the NAO robot via a gesture recognition system.

1.1.1 Recognition of gestures

The analysis and automatic recognition of human movement from visual input is one of the most treated areas of research in computer vision. This is not only due to its exciting scientific challenges, but also because of its practical importance. Hundreds of existing potential applications in urgent need of this technology include control, navigation and manipulation in real and virtual environments, human-robot interaction, tele-presence

systems, clinical diagnosis and monitoring, assistance to the elderly, learning applications and games, engineering systems and computer aided design, automatic annotation and indexing of videos, forensic identification and lie detection, etc. As part of our project where our task is to endow the humanoid robot NAO with the ability to recognize movements of people, we address the problem of recognizing human actions from video streams. The latter has interesting advantages such as free noise and low cost. In addition, the video has semantically very rich and relevant content for the recognition of human actions. In this context, we address the subject of expressive gestures in order to make communication more effective and give it a high level of impact.

1.1.2 Human-Robot interaction

Robots are mainly used in industry, but with the development of technology they have become more and more introduced into our daily life. These robots, intended to be put in direct contact with humans, must be able to interact intuitively in order to be accepted by humans. This implies that the robot must be able to perceive its surroundings and act accordingly in a humanly acceptable and welcoming manner. Traditionally, autonomous robots have been employed in areas where little or no interaction is required, such as; the automotive industry, inspection of oil wells, research and exploration of environments hostile to humans. These robots are generally remotely controlled and supervised by a human operator. Recent additions of service robots to home environments, such as robot vacuums, have increased their contact with a common person, however, no high-level interaction is involved with humans. Service robots are now evolving to acquire more important roles, such as assistants, hospital workers or caregivers to the elderly, where social interaction is an important part of robot activity. The presence of such robots in our daily life is important for some people, particularly those with reduced mobility, who require an assistant capable of acquiring and interpreting information and thus transmitting it. Today, we find a number of projects carried out to ensure a social interaction between man and robot, we can cite [237], carried out in the Institute of Technology of Karlsruhe. This project consists of a recognition of the user's speech and a visual perception of his pointing gestures and the orientation of the head. The combination of both video and audio channels allows multimodal speech analysis that leads to higher level interpretation, such as user intention analysis in a robot dialogue situation. We also cite the HUMAVIPS [73] project carried out at INRIA which consists in providing the NAO robot with audiovisual (AV) capabilities: exploration, recognition and interaction, so that it exhibits adequate behavior when it comes to a group of people. There is also the ROMEO [203] project which consists in creating a humanoid robot companion and personal assistant.

1.2 Motivation

A widely held assumption in the field of human-robot interaction (IHR) is that a good interaction with a robot should reflect the most natural communication possible. For this we always seek to draw inspiration from human-human communication which requires words and gestures as explicit modalities. But in communication between humans there is also another channel, implicit, through which information about individuals is transmitted, usually in the form of emotions. This implicit channel is an essential characteristic of human communication.

Therefore, if we want to obtain a truly effective interaction of robots with humans, we must also address this type of communication. So, the objective of our work is to develop a system for recognizing human gestures, while considering their movements and also their moods. Although the video content is quite informative, the task of recognizing human action remains problematic. Indeed, large variations of style can be observed in the reproduction of the same action depending on several factors. In addition, interclass ambiguity must be taken into account since there are similar actions such as the actions "run slowly" (jogging) and "walk". In our work, in the first section we are inspired by Laban's movement analysis model (LMA) for the representation of human movements. This model has been used in several applications for different research purposes, in dance [14], music [251], robotics [159, 131, 129, 126, 174], etc. In our work, we rely on the components of this model (Body, Space, Form and Effort) to describe the quantitative and qualitative aspect of the gesture. And in the second part we try to build our descriptor by inspiring some geometric and time depended features [148, 296, 297, 185, 92]. Since we tried to create a platform that, in addition to facilitating the recognition of human movements for the robot, also helps to recognize the movements of a robot for another robot, to make this descriptor, in addition to human movements to perform various gestures, We also noticed to the movements of the NAO robot.

1.3 Contributions and publications

1.3.1 Contributions

Our work in this thesis aims to construct a gesture recognition system that is adapted to real conditions. It includes the following contributions:

- In the first part, we created a strong descriptor using the LMA method. Since our goal in this section was simply to recognize human movements regardless of their moods, only three of the four LMA factors were used to construct this descriptor. Then, using the dynamic time warping algorithm, we updated this descriptor so that we can select the most important factors that play a role in creating a suitable pattern for identifying movements. The output of the dynamic time warping algorithm are curves that have noise due to various conditions such as speed and acceleration. For this reason, we eliminated the noise using the Fourier transform, and finally evaluated the accuracy of the proposed system using support vector machine [199].

- we also used a new method based on the geometric coordinates of different parts of the body to present a movement. To do this, in addition to the distances between hip center and other joints of the body and the changes of the quaternion angles in time, we defined the triangles formed by the different parts of the body and calculated their areas. We also calculated the area of the single conforming 3-D boundary around all the joints of the body. At the end we added the velocity of different joint in the proposed descriptor. We used a multi-layer long short-term memory (LSTM) network to evaluate this descriptor. The proposed algorithm has been implemented on five public data sets, MSRAction, UTKinect, Florence3D, SYSU 3D HOI and NTU RGB+D 120. The results are compared with those available in the literature.

- We have developed a system for recognizing expressive gestures. The latter is made up of a data set

made up of 5 expressive gestures, interpreted with 4 emotions (joy, anger, sadness and neutral). The local movement descriptor presented in the previous chapter was used and modified according to the global measures to describe the entire expressive gesture. With the integration of the Effort component, responsible for describing the expressiveness of the movement, we create an overall expressive movement descriptor. The system is evaluated with public data sets and our expressive data set.[7]

- In order to improve the identification of movement and emotions in the data set described in the previous item, we created a descriptor using temporal and spatial factors. We also implemented a feed forward neural network to evaluate the new descriptor. Using this recognition system, the accuracy of motion detection in this data set reached 100% and also the percentage of accuracy of emotion recognition was significantly increased.

1.3.2 Publication

- AJILI, Insaf, RAMEZANPANAH, Zahra, MALLEM, Malik, et al. Expressive motions recognition and analysis with learning and statistical methods. Multimedia Tools and Applications, 2019, vol. 78, no 12, p. 16575-16600 [7].

- RAMEZANPANAH, Zahra, MALLEM, Malik, et DAVESNE, Frédéric. Human Action Recognition Using Laban Movement Analysis and Dynamic Time Warping. In : 24th International Conference on Knowledge-Based and Intelligent Information Engineering Systems (KES 2020) [199].

- RAMEZANPANAH, Zahra, MALLEM, Malik, et DAVESNE, Frédéric. Emotion Recognition Based On Dynamic BodyMovement.Journal Multimedia Tools and Applications, 2020. (Under Review)

- RAMEZANPANAH, Zahra, MALLEM, Malik, et DAVESNE, Frédéric. Autonomous Gesture Recognition using Multi-layer LSTM Networks and Laban Movement Analysis. International Journal of Knowledge-Based and Intelligent Engineering Systems, 2020. (Under Review)

1.4 Organization of the thesis

- Chapter 1 presents the general context, the motivation of this work, and the contributions with the various publications produced in these 3 years.

- Chapter 2 provides a state of the art divided into 2 main parts: the first part focuses on body language where we presented the modules of a gesture recognition system and the proposed movement descriptors. The second part is based on emotional language, where we defined the theme "emotion" and the various methods proposed in the state of the art to express emotion, and a final part on expressive gestures.

- Chapter 3 deals with the identification of human movements. For this purpose, a local descriptor based on the Laban Movement Analysis, LMA, is defined to describe the movements. LMA is an algorithm to de-scribe a motion by using its four components: Body, Space, Shape and Effort. Since the goal in this part is gesture recognition, only the first three factors have been used. The Dynamic Time Warping,

DTW, algorithm is implemented to find the similarities of the curves obtained from the descriptor vectors obtained by the LMA method.Finally, the Support Vector Machine, SVM, algorithm is used to train and classify the data. The proposed pipeline is evaluated on public data sets.

- Chapter 4 presents the descriptor which is based on the geometric coordinates of different parts of the Skeleton which can be considered a subset of body components of LMA. We also added some time-dependent characters to this descriptor. According to the definition of "Effort" in LMA, all time-dependent elements can be considered as a subset for the Effort components. The long short-term memory (LSTM) network is used to evaluate this descriptor. The proposed algorithm isimplemented on two public data sets.

- Chapter 5 presents a system for recognizing expressive gestures. Two expressive motion descriptor based on LMA are constructed. A comparative evaluation between 2 learning methods (Random Decision Forest and Feed Forward Neural Network), and a human approach method is carried out. The model is evaluated on public data set and our own data set made of expressive gestures.

Chapter 2

State Of The Art On Body Gesture And Emotion Recognition

Contents

The problem of gesture recognition usually involves describing the content of the sequence in order to understand what is happening in a video file. This high level information allows later to identify a video in a data-set. So the problem is to extract the characters associated with a movement over time, called the descriptor motion. There are two types of descriptor, low-level descriptors based on the raw content of the image (color,

points of interest, texture, etc.) and high-level descriptors oriented to the semantics of the content of the scene for describe the messages transferred by the gestures or the emotional state expressed by the people during the realization of the movement. This chapter has three main parts. The first chapter, defines the principle of a gesture recognition system with its different modules. We present the different motion descriptors used in the framework of the recognition of the actions as well as the public data set of actions. The second part is about the emotional language where the different modalities used for the analysis and recognition of emotions are defined. Subsequently, we focus on the aspect of expressive gestures that will be our topic of thesis. In the third part, we introduce the motion analysis model (LMA) that we used in our work to interpret and describe expressive gestures.

2.1 Sign language

2.1.1 Definition of a gesture

A gesture can be defined as the elemental movement of a person's body parts. It is part of the nonverbal communication used instead or in combination with a verbal communication, to express a particular message. Interpreting gestures is a complex task, mainly because gestural meaning involves a cultural context. For example, a nod usually indicates an agreement, but in some countries (such as Greece) a single nod indicates a refusal. Scoring with a thumbs up gesture is a common gesture in the United States and Europe, but it is considered a rude and offensive gesture in Asia. In Cistercian monks, when a person approaches his finger next to his eye and then directs it to the eye of his interlocutor, or if he touches the heart and then directs his hand towards the other, this is a sign of suspicion. In India, however, these signs signify peace. All this implies that the semantic interpretation of a gesture depends strictly on the given culture. The gesture can also be classified into a dynamic and static gesture. The latter is also called posture, corresponds to the configuration of the body or part of the body at a given moment [227] while the dynamic gesture [24] corresponds to a continuous succession postures. As the literature shows, this is not the only way to classify gestures:
In Cadoz's classification [43], the gestures are divided according to the gestural functions into 3 groups:

- Epistemic gestures: gestures for perception with touch.
- Ergotic gestures: gestures act on the world, can manipulate, modify or transform the physical world.
- Semiotic gestures: gestures of expression to communicate information and send messages to the environment.

In Kendon's classification [120] gestures are classified as a continuum that distinguishes between gestures accompanying speech (referred to as gesticulation) and those that are independent (autonomous).

In McNeill's classification [163], based on the Kendon continuum, gestures are divided into 4 groups:

- Iconic Gestures: gestures that illustrate concrete concepts.
- Metaphorical gestures: gestures represent abstract concepts and metaphors.
- Deictic gestures: pointing gestures toward a referent.

- Beats: gestures rhythmic speech accentuating the important elements, without semantic content.

From the following definitions, we place our work in the context of gesture recognition with semantic content. Specifically we will talk about expressive gestures, that means gestures made with different emotions.

2.1.2 Gesture recognition system

The recognition of human gestures is a process of automatically identifying and interpreting human movements. Three basic steps in a gesture recognition system are:

- The acquisition of data which consists of extracting digital information through a motion capture system.
- Feature extraction to convert raw data into a meaningful representation of motion.

  The goal of this step is to extract the useful and compact features that represent the movements in the most reliable way possible. These characteristics constitute a specific descriptor vector for each gesture.
- The learning model, in which the vectors obtained from the previous steps are used to train and classify.

In the following, we will have a brief explanation of these items.

2.2 Motion capture devices

The first sensors in the field of vision were based on 2D RGB images [218]. However, they only provide the appearance information of the objects in the scene. With this limited information, it is extremely difficult, if not impossible, to solve some problems such as foreground and background separation with similar colors and textures. As a result, recent studies tend to use new information that is depth. This information solves the problem of 3D inference. For this, in recent years, this area has experienced a strong presence of 3D sensors that have improved the performance of gesture recognition. There are several techniques for capturing the depth of a scene. Some are based on the principle of matching which consists of finding the points that correspond between the right and left images, Figure 4.4. In the following, we will review the types of sensors. The popularity of these type of sensors can be explained by the analogy with the human visual system and

Figure 2.1: Xovis 3D Stereo

the availability of low cost color cameras. The authors in [233], used this type of sensor for gesture tracking. They detected the user's fingertip in both images of this stereo pair. In these images the two points on which

this end appears establish a stereo match, which is used to evaluate the position of the finger in the 3D space. This position is used by the system to estimate the distance of the finger from the augmented table and hence determine whether the user is in contact with it or not. This stereoscopic information was also used by [107] to track and recognize the gestures of both hands, while solving the problem of horizontal noise generated by the camera because of its hypersensitivity to light. However, stereoscopic vision systems can only calculate the depth of the scene for a restricted set of pixels corresponding to areas of the scene having a strong local structure. For this purpose, a new 3D camera family has appeared, the flight time camera (TOF) shown in Figure 4.5. It is an active sensor that provides depth images in real time based on the measurement of flight

Figure 2.2: 3D Time-of-Flight (TOF) Cameras

time. Its principle is similar to that of a laser scanner. It consists in illuminating the scene and the objects measured by a flash of light, and calculating the time that this flash takes to make the flight between the object and the camera. The flight time of this flash is directly proportional to the distance between the camera and the measured object. This measurement of flight time is performed independently by each pixel of the camera, thus obtaining a complete 3D image of the measured object. In this context, [98] used the SwissRanger camera for the acquisition of arm gestures. The motion was detected by the difference between two image ranges and was filtered by a band-pass filter. Their motion descriptor was based on the shape context to ensure invariance of the rotation. The correlation between the different representations of the contexts of forms was carried out in the gesture recognition phase. Similarly, [37] used the TOF camera to acquire hand gestures. The data is transformed into point clouds after a noise filtering phase.The principal component analysis (PCA) was applied to obtain a first estimate of the position and orientation of the hand. A matching principle has been realized to minimize the distance between the model and the point cloud. [67] used this sensor for pointing recognition. They extracted three type of information for the representation of the gestures (distance between the head and the hand, the angle between the arm and the vertical axis of the body and the speed of the hand). They applied the Gaussian Process Regression Method (GPR [200]) to model an approximation of a function that associates extracted body characteristics to a pointing direction. The advantage of the TOF camera is the speed of image acquisition, as each pixel of the camera independently delivers a measure of distance. In return, this type of camera generates a cloud with a large number of points. This may require large storage space on used equipment. In addition, physical limitations related to sensor size can result in low resolution depth images at a small range or at significant measurement noise. Hence the appearance of a second depth sensor which is the kinect, this camera as the TOF provides the depth of image but with a different resolution. In fact, the TOF cameras still have a limited resolution (200×200) while the kinect has a VGA resolution (640×480). RGB-D data provides a very useful representation of an indoor scene to solve fundamental problems in computer vision. The Kinect sensor takes the benefits of the color image that provides appearance information of an object and

the depth image that is immune to variations in color, lighting, angle of rotation and scale. However, with the output of this low-cost sensor, acquiring RGB-D data becomes cheaper and easier.

2.2.1 Kinect depth sensor

The Kinect is a very competent evolution of a classic camera coupled with an Xbox 360 allowing the user to control video games without the use of joysticks, but only with the movements of the body. It was designed by Microsoft in September 2008. It is a device suitable for 3D reconstruction of indoor scenes. It incorporates three components (See Figure 2.3):

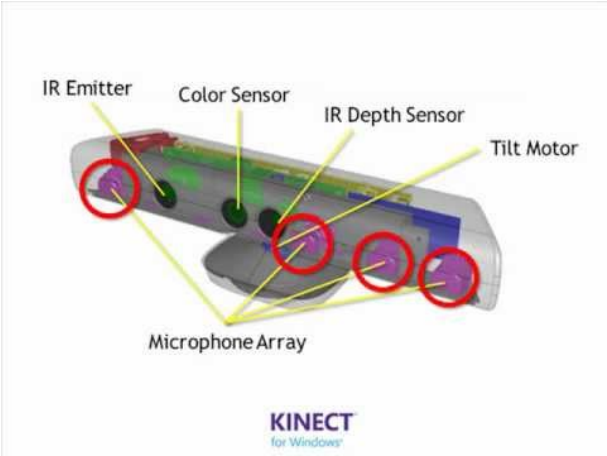


Figure 2.3: kinect for windows

- A microphone for voice control allowing the functionality of voice recognition.
- A RGB camera providing a color image with a resolution of 640×480 pixels at an average frequency of 30 frames per second.
- A 3D depth sensor: An infrared emitter (laser) and an infrared camera to calculate the depth of an object relative to the camera.

2.2.1.1 Kinect-specific tools

OpenNI is an open source framework capable of interfacing with different types of hardware and designing applications for natural interactions. This Framework is accompanied by an open source "libfreenect" for Kinect. This driver provides all the basic functions of Kinect (depth, Infrared, RGB) and allows this sensor to be used with Linux and ROS. The processing algorithms available under this framework are:

- Body analysis: recovery of the skeleton pose by determining the position and orientation of its joints.
- Analysis of the scene: extraction of the foreground, segmentation of the ground, etc.
- Gesture detection and monitoring of hand movements.

In addition to the advantages of Kinect sensors, its disadvantages should also be mentioned. One of its disadvantages is the noise is so high that in many cases, for example, a chair was recognized as a human being. Therefore, this type of sensor can be very accurate only in cases where we are indoors, and even if a hall is

empty of other objects. For this reason, wearable sensors are another sensor that has been widely used by researchers. In the following, we will explain this type of sensors and introduce the sensors used in this thesis.

2.2.2 Inertial sensors

Wearable motion sensors are the most popular and common wearable devices for receiving information and analyzing movements and physical activities related to daily life. Their structure is basically a combination of an accelerometer and a special gyroscopic sensor, to evaluate the movement, and their small size makes them easy to wear on different parts of the body.

2.2.2.1 Principles of Inertial Sensors

To obtain information such as three-dimensional location, speed and acceleration, a coordinate system must be defined before installing an inertia sensor. In general, by considering (A_0) as the acceleration of the origin of a moving coordinate system which rotates with a gyroscopic angular velocity (w), and if a mass (m) has the position vector (r_0) and the velocity (u_0) with respect to the moving coordinate system, then the observed inertia force, [242], is as follows: :

$$mA' = -mA_0 + 2mu' \times w + mw \times (r' \times w) + mr' \times \frac{dw}{dt} \tag{2.1}$$

The right of the equation are four variables derived from angular accelerations related to linear inertia, Coriolis, centrifugal forces, and apparent forces. As a result, the analysis of the moving coordinate system has complex calculations and and signal handling must be taken into account.

Accelerometers

The acceleration of an object is the first derivative of velocity or the second derivative of its displacement. However, deriving signals increases the noise. Therefore, direct acceleration measurements are often simpler, easier, and more reliable. According to Newton's second law, the acceleration of linear motion, α is the force (F) acting on a mass (m), i.e:

$$F = m\alpha \tag{2.2}$$

Many accelerometers are available with different specifications and prices. Accelerometers, for example, are relatively inexpensive and reliable for use as shock sensors in cars. Novel MEMS technology is used to produce small and sensitive accelerometers. In general, the correct type of accelerometer should be selected for each specific application. Beam type accelerometer is the most sensitive type of accelerometer in calculating the acceleration of body motion. In this type of accelerometer, an elastic beam is fixed to the base at one end, and a mass called a seismic mass is attached to the other end as shown in Figure 2.4. When a seismic mass accelerates, a force proportional to the mass of acceleration is applied to it, and the beam bends in response to the force. Adequate damping coefficient must be established in the mechanical system to prevent resonance oscillation after transient input. Instead of a beam, a diaphragm, spring, or any other elastic material can be

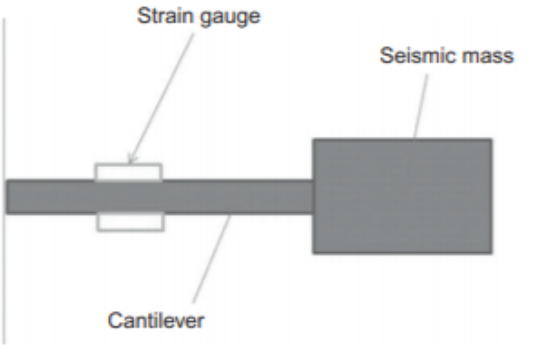


Figure 2.4: Uniaxial accelerometer

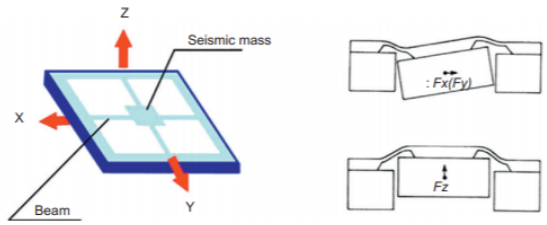


Figure 2.5: Tri-axial accelerometer.

used in the accelerometer. Estimating the amplitude and direction of acceleration in 3D requires a three-axis accelerometer, Figure 2.5. MEMS technology-based three-axis accelerometers are commercially available. The displacement of a moving object can be measured using a variety of methods, including those based on the effects of a piezoelectric, piezoelectric, or capacitor. Often, a barometer or accelerometer-displacement semiconductor is used to measure motion in humans and animals because it is small and relatively inexpensive. In general, a piezometer element is a pressure gauge attached to or embedded in a mass-loaded concealer beam. As the beam bends in response to acceleration, the resistance changes. Four separate measuring elements are arranged in the Wheatstone Bridge configuration, which calculates the output while eliminating cross-axis, temperature and other transparent inputs. Using micromashing technology (MEMs), piezorestores can be easily implanted in support beams that attach seismic mass to the frame or backing. Piezoelectric accelerometers are usually used when only different acceleration components need to be measured. Very low power consumption, simple detection circuits, high sensitivity and inherent temperature stability characterize piezoelectric accelerometers. A polar voltage occurs in the piezoelectric material that is proportional to its deformation. The polarity of the polarization voltage depends on the molecular structure of the material. Figure 2.6 shows an example of a bimorph configuration beam that has two piezoelectric elements with different poles that produce a double or different output. A three-axis biomorph has also been developed [302]. The terminal voltage of the piezoelectric sensor is due to the load, which is caused by the bending of the piezoelectric element and its capacitor. The terminal voltage is proportional to the stored load. More accurately, a load amplifier can be used to measure the load generated. While the capacitance of the input, which includes the capacitance of the piezoelectric element , C_f, and the capacitance of the stray, is denoted by C_d, we have the following equations:

$$Q = C_d V_i + C_f (V_i - v) \tag{2.3}$$

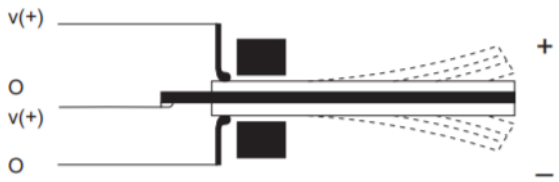


Figure 2.6: Configuration of a bimorph.

And

$$V = -AV_i \tag{2.4}$$

where Q is the generated charge and V_i, V, and A are the input voltage, output voltage, and the gain of the amplifier, respectively. If $A >> 1$ and $AC >> C_i$, then

$$V = \frac{-Q}{C} \tag{2.5}$$

Where C is the input capacitor. Acceleration signals are used for a wide range of measurements, including assessing balance, shifting walking and sitting, classifying movements, reducing physical activity, and estimating the cost of metabolic energy.

Gyroscopic Sensors

The gyroscope is responsible for measuring angular velocity. The sensor consists of a rotating wheel mounted on a movable frame mounted. When the sensor is rotating, it tries to maintain its original orientation in space, even with the central forces acting on it. When the direction of the axis changes externally, a torque is created that is proportional to the rotational speed of the inclined axis, which can be used to detect angular velocity. An example of such a converter is a dynamically tuned gyroscope. Gyroscopes are the same vibration accelerometers that measure Coriolis forces (Figure 2.7). A conventional vibrating gyroscope consists of an anti-mount object mounted on a suspension that allows the anti-motion object to move in two orthogonal directions. To produce the Coriolis force, the proving mass must be in motion. For this purpose, the anti-electron mass is forced to oscillate in a direction parallel to the surface of the chip. If the direction of rotation of the gyroscope chip is around an axis perpendicular to the chip surface, the Coriolis force causes the proving mass to deviate in the second direction.

2.2.2.2 Xsens MVN Awinda [213]

Xsens MVN, which consists of hardware and software, is used to capture human movements using wearable sensors.

The hardware of Xsens MVN 2.8 consists of wearable sensors that streams data wirelessly to a PC/laptop (MVN Link) through completely wireless solution (MVN Awinda). To capture the motion of the human movement, 17 motion trackers are attached to the feet, lower legs, upper legs, pelvis, shoulders, sternum, head, upper

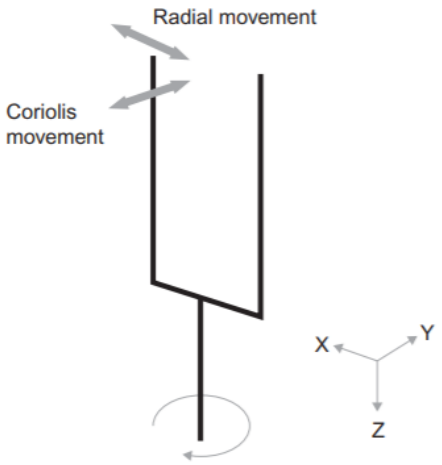


Figure 2.7: The Coriolis force.

arms, fore arms, and hands. Sensor modules are units of inertial and magnetic measurements that, in addition to containing 3D gyroscopes, 3D accelerometers, and 3D magnetometers, have an advanced signal processing pipeline that includes StrapDown Integration (SDI) algorithms in order to send the data at a proportionately low rate (60Hz), while maintaining the accuracy of sampling at a much higher rate (e.g. >1kHz). Thanks to the SDI, 3D tracking accuracy of each sensors is equivalent both for MVN Link (240Hz) as well as MVN Awinda (60Hz), with only a reduced time resolution for the latter. For each participant, we collected MVNX files. It is an XML format that can be imported to other software, including MATLAB and Excel. This format contains several information (Table 2.1), including sensor data, segment kinematics, and joint angles, as well as subject information needed to recreate a 3D visualization of an avatar.

Data	Definition
Sample Counter	To show that data from all MTw's are correctly allocated.
SDI data	Velocity Increment, Orientation Increment
Inertial and magnetic	3D acceleration, angular velocity, magnetic field and pressure
Orientation data	Quaternions, Euler angles, Rotation Matrix (Direction Cosine Matrix)
Awinda wireless network properties	Received Signal Strength Indicator
Status Word	A 32-bit output of ones and zeros indicating the status of filter and components
Clipping flags	indicating if a given sensor component has exceeded its range

Table 2.1: Saved and Exported Data

2.3 Features extraction

Recognition of human action is one of the major subjects in the field of computer vision. This is due to the wide variety of potential applications, such as video surveillance, video content analysis, sports training, health monitoring, business behavior analysis, human-machine interaction, robotics, etc. However, this process is experiencing real difficulties due to the high variability of people in appearance and in movement. Therefore, it is essential to extract robust representations from these variations. Several types of descriptors have been proposed in the computer vision literature. We have chosen to subdivide them into two types: descriptors based on local characteristics (points of interest) and descriptors based on appearance (3D silhouette and skeleton).

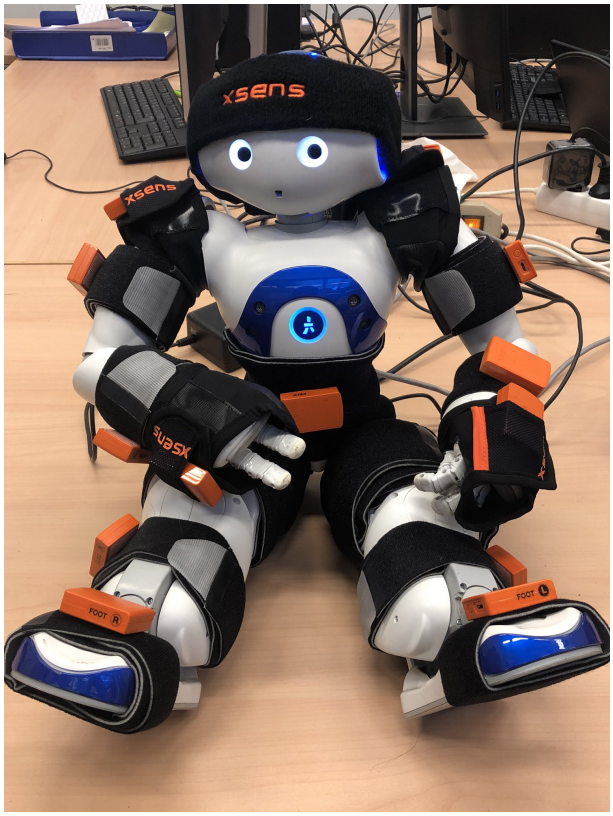


Figure 2.8: The Coriolis force.

2.3.1 Methods based on local characteristics

The first approaches in the state of the art used methods known in the field of image recognition. These methods are based on local characteristics which generally correspond to a set of pixels having a singularity, either at the level of the gradient or the contour. These methods use point of interest detectors in images to represent them as collections of items of interest. Spatio-temporal points of interest (STIP) are defined as the points in the image where a significant change in time and space appears. It is an extension of the spatial points of interest (noted SIP for "Spatial Interest Points"). They are identified from the local maxima of a response function which characterizes the Spatio-temporal signal. One of the first works proposed to extract spatio-temporal points of interest (STIPs) is that of [135].They proposed the Harris 3D point of interest detector (See Figure 4.6), which has a temporal extension of the Harris2D wedge detector [93] to recognize human actions. It detects points whose local neighborhood is subject to significant spatial and temporal variation.

However, the number of points of interest satisfying the Harris3D criterion is relatively small compared to the

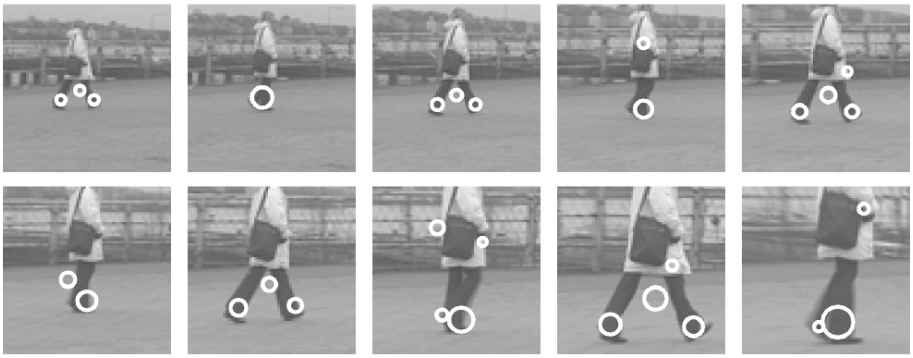


Figure 2.9: 3D corner detector from [135].

zones containing significant movements. [66] have addressed this problem with a new detector specially designed for local periodic movements present in the video (see Figure 4.9). They proposed new denser spatio-temporal points of interest. Their detector is based on 2D Gaussian filters applied to the spatial dimension and Gabor 1D filters applied to the temporal dimension.

- The Gaussian filter performs a spatial scale selection (σ) by smoothing each image.
- The Gabor band-pass filter gives high responses to periodic variations of the signal.

The authors concatenated the calculated gradients for each pixel in a cuboid region into a single vector. Finally, the Principal Component Analysis (PCA) method was applied for the projection of vectors over a smaller dimension space. This detector, named Cuboid, was applied initially for the recognition of the movements of an animal and later for the actions and the facial expressions of the people.

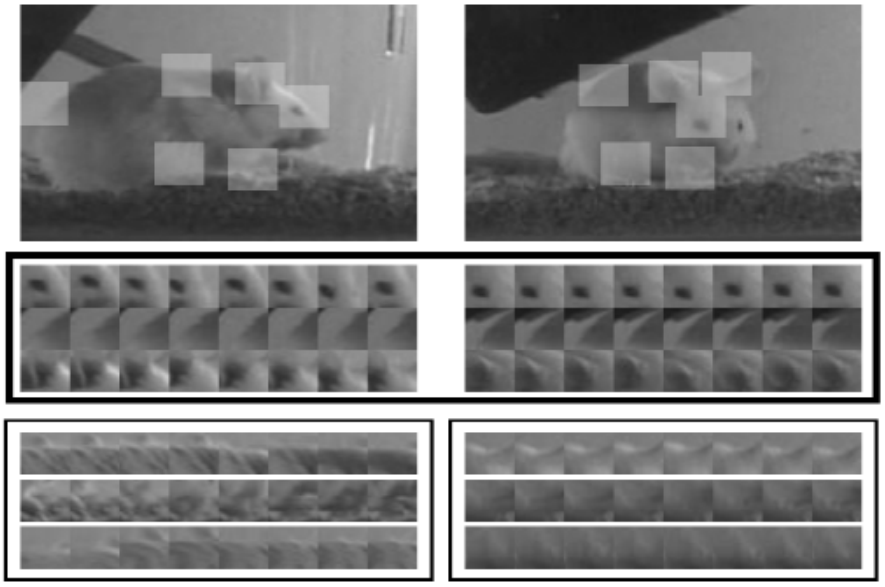


Figure 2.10: Cuboid detector of [66].

In [230], the authors proposed an extension of Motion History Image (MHI) in order to map the frames that make up a gesture. To do this, they first separated the background from the foreground using a median filter. They then extracted the participant information using edge detection. In this way, a video was considered as a set of frames containing information obtained from the edge detection algorithm. Finally, they projected this three-dimensional spatio-temporal information onto the planes of three different view, XY, YT, XT in which X, Y and T represent height, width and time respectively (Fig 2.11).

Following [230], the authors in [208], proposed the Bag of histogram of optical flow (BoHOF) in order to represent the velocity of each gesture. To do this, they extracted histograms of oriented gradients (HOG) feature, proposed in [61], from the three Motion History Image (MHI) planes, I_{XY}, I_{YT}, I_{XT}. Then they divided the body into four categories, the right hand, the left hand and the right and left legs were the four parts they suggested. Finally, for all body parts, the difference between a similar pixel in two consecutive frames was calculated. The authors of [128] proposed a new local descriptor for video sequences, the HOG3D. Which is an extension of the (HOG) in the space-time domain. Gradients are calculated using a full video representation. Regular polyhedra are used to uniformly quantify the orientation of the space-time gradients. Their descriptor combines shape

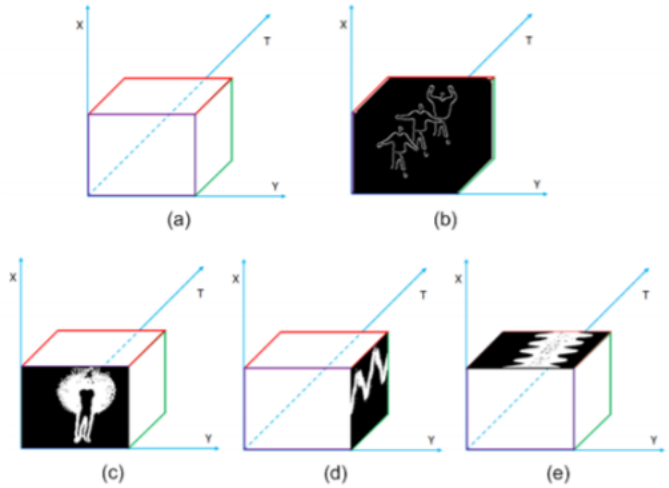


Figure 2.11: The proposed representation of a video in [230]. The figure (a) shows the axis formation of 3D planes (b) modeled sequence of frames in 3D volume space (c) projection on XY plane (d) projection on XT plane (e) projection on Y T plane .

and movement information at the same time. Figure 2.12 shows the principle of calculation of the HOG3D descriptor. The researches in [136] also used HOG oriented gradient histograms combined with optical flow

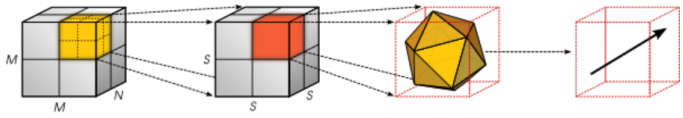


Figure 2.12: A Spatio-Temporal Descriptor based on 3D Gradients (HOG3D) [128]

histograms (HOF) in order to characterize local movement and appearance. In this regards, they calculated the descriptors' histogram of the space-time volumes in the vicinity of the detected points. Each volume is subdivided into a grid of cells $n_x \times n_y \times n_t$. For each cell, the histogram with 4 components of gradient orientations and 5 components of optical flow orientation are calculated. The normalized histograms are concatenated in the final descriptors HOG and HOF. In [276], the authors extended the Hessian 2D characteristic in the space-time domain by applying a 3D Gaussian filter. They calculated the determinant of the Hessian matrix in order to determine the points of interest. Chakraborty et al.[51], proposed a selective detector of spatio-temporal points of interest for the recognition of actions human. Local space-time constraints are imposed to obtain a more robust final set of points of interest, while removing unwanted points of interest. The support vector machine (SVM) method was used for classification and recognition of actions. Yan and Luo [283], proposed the histogram of interest point locations (HIPLs) algorithm as a complement to the descriptor of the bag of interest points (BIPs) in order to capture the spatial–temporal interest points (STIPs) information. In their approach, the Adaboost boosting method, [78] and the sparse representation (SR) were used with the weighted output classifier (WOC) to achieve a better classification of all the characteristics. AdaBoost is an algorithm for boosting which is based on the iterative selection of a weak classifier according to a distribution of learning examples. Each example is weighted according to its difficulty with the current classifier. This algorithm improves the performance of the classification model. The sparse representation was applied in order to manage

the large intra-class variations of human actions. WOC is a framework for merging multiple characteristics which consists in exploiting the potential of each characteristic and using weights to combine weak classifiers driven by a unique type of characteristic. However, the HIPLs model does not provide any temporal information about the data in the video, such as speed. For this, this system is unable to correctly discriminate the action classes which are very close to each other, such as running and jogging. Cao et al. [48], addressed the problem of detecting actions from cluttered videos. For the detection of actions, they combined multiple characteristics, which can be based on movements (for example, history of movements, optical flow) or on appearance (for example, edge, color). In addition to this, they used heterogeneous characteristics such as the hierarchical filtered motion field (HFM) [248], scattered characteristics [66], histograms of oriented gradient and optical flow (HOG / HOF) [136]. They used the Gaussian mixture models (GMM) to model and combine these heterogeneous characteristics. The probability of a given characteristic vector is estimated based on the GMM model. The authors in [179], presented a new activity recognition descriptor from videos acquired by a depth sensor. This descriptor presents a histogram capturing the distribution of the normal orientation of the surface in 4D space, time, depth, and spatial coordinates. As shown in Figure 2.13, for the construction of this histogram, they created 4D projectors, which quantify the 4D space and represent the possible directions for the 4D normal. Projectors are initialized using the vertices of a regular poly chore. Therefore, quantization is bypassed using a new discriminating density measure, so that additional spotlights are induced in directions are more dense and discriminating. Finally, the SVM method was adopted for training and classifying actions. [ang et al. [288],

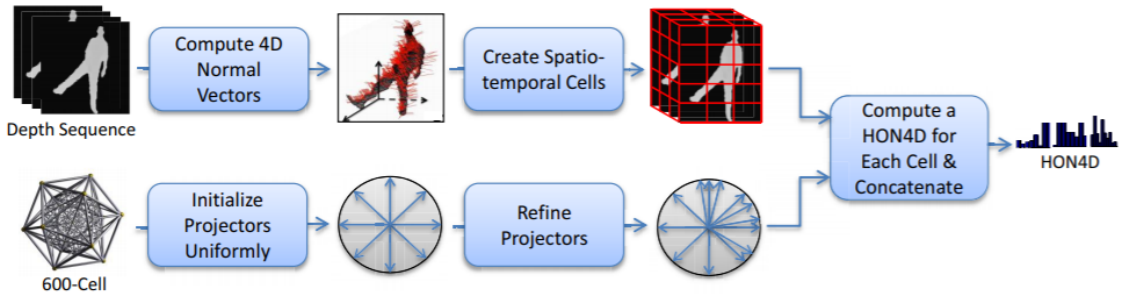


Figure 2.13: The various steps for computing HON4D descriptor.

used oriented gradient histograms (HOG) calculated from depth of motion maps (DMM), as a representation of an action sequence. They projected each depth map onto three redefined orthogonal planes. Each projected map was normalized and a binary map was generated showing its movement energy by calculating the difference between two consecutive maps. The binary maps are then stacked to obtain the DMM for each projective view. The histogram of oriented gradients is then applied to the DMM card to extract the characteristics of each view. The concatenation of the HOG descriptors from the three views forms the set of DMM-HOG descriptors which present the inputs of a linear classifier SVM for the recognition of actions. An illustration of the steps for extracting HOG from DMM is presented in Figure 2.14.

2.3.2 Appearance-based methods:

The following is a review of research that used a ghost or human skeleton to describe a gesture.

- Silhouette-based methods: These methods use silhouettes as input for the human action recognition

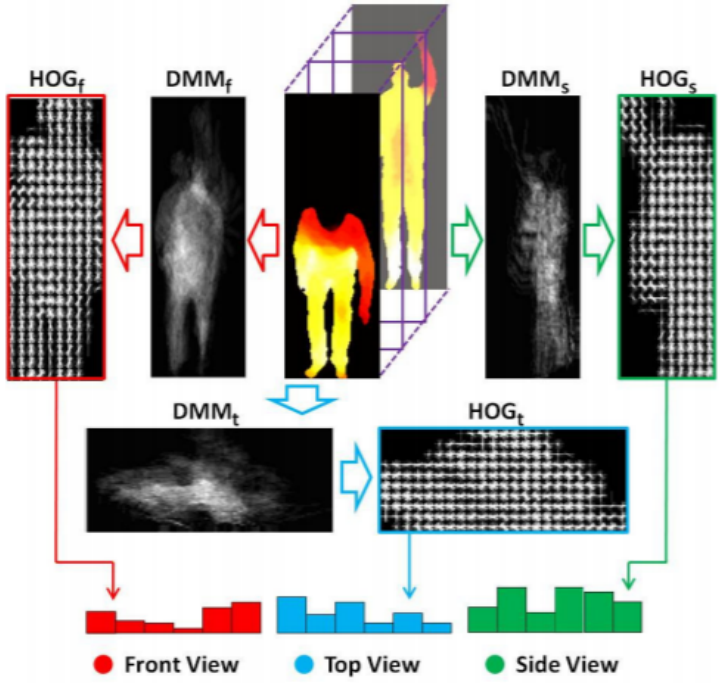


Figure 2.14: The framework of computing DMM-HOG. HOG descriptors extracted from depth motion map of each projection view are combined as DMM-HOG, which is used to represent the entire action video sequences.

system. A 3D space-time volume (STV) is formed by stacking images during a given sequence. Precise localization, alignment and possibly subtraction of a background are necessary. Weinland et al. [274], introduced new motion descriptors (MHV) based on motion history volumes which is an extension of MHI in 3D [33]. They transformed the calculated MHVs into cylindrical coordinates around the vertical axis and extracted invariant characteristics in the Fourier space. In order to evaluate their method, the authors built a database, called IXMAS, composed of the actions captured from different points of view. The results on this basis indicate that this representation can be used to learn and recognize the classes of actions regardless of gender, body size and point of view. The same authors in [273], proposed using a set of representative silhouette as models. In their work; actions are represented by distance vectors between the instances and the images in the action sequence. In addition, different methods of selecting key pose silhouettes were discussed in their work. Ahmad and Lee [3], proposed a spatio-temporal representation of silhouette, called silhouette energy image (SEI) to characterize the properties of form and movement for the recognition of human action. To address the variability in the recognition of actions, they proposed adaptable models with several parameters, notably the anthropometry of the person, the speed of the action, the phase (initial and final state of an action), the observations of the camera (distance from the camera, oblique movement and rotation of the human body) and variations in views. The authors in [72], first reduced the size of the silhouettes to small points as a description of the spatial movement using the Locality Preserving Projections (LPP). This motion vector obtained after reduction of dimension was taken to describe the structure of intrinsic motion. Then, three different temporal information, the temporal neighbor, the difference in movement and the movement trajectory, were applied to the spatial descriptors to obtain the characteristic vectors, which were the inputs of the classifier k nearest neighbors (KNN). The same authors in [252], have developed another approach based on the silhouette. They used the locality adaptive preserving projections (LAPP) in order to construct a discriminating spatio-temporal subspace.

Then, the method called Non base Central-Difference Action Vector (NCDAV) was used to extract the temporal data from the reduced spatial subspace in order to characterize the movement information in a temporal vector. This makes it possible to solve the problem of overlaps in the spatial subspace resulting from the ambiguity of the shape of the human body between different classes of action. Finally, the method of learning a large margin nearest neighbor (LMNN) was applied to construct a discriminating space-time subspace where the time vectors belonging to the same action class are grouped together while those associated with different classes are separated by a margin. In the action classification step, the authors used the approach of k nearest neighbors. However, their solution depends on the quality of the extracted silhouettes, which makes the recognition stage more sensitive. In [89], the authors considered an action as a temporal sequence of local shape deformations of the silhouette object centroid. Each action is represented by a covariance matrix of the vectors of geometric characteristics normalized in 13 dimensions which capture the shape of the silhouette tunnel. The silhouette tunnel of a test video is divided into short overlapping segments and each segment is classified using a dictionary of labeled action covariance matrices and the nearest neighbor rule. In [125], the researchers proposed the concept of the accumulated motion image (AMI) to represent the spatio-temporal characteristics of actions. The AMI was presented according to the differences between to consecutive frames by:

$$AMI(x,y) = \frac{1}{T}\Sigma_{t=1}^{T}|D(x,y,t)| \tag{2.6}$$

Where $D(x,y,t) = I(x,y,t) - I(x,y,t-1)$ and T donates the total number of frames. By calculating the distances from the query action movie ranking matrix to the ranking matrices of all local windows in the query movie, local windows close to the query action are identified as candidates. To find the best fit among the candidates, their energy histograms, obtained by projecting AMI values horizontally and vertically, respectively, are compared with what is presented in the query video. Figure 2.15 shows an example of someone sitting from a data set used in [32]. In [225], the authors defined their descriptor using silhouette for the recognition of the actions. They used the body pose histogram (BPH) to describe a sequence of human action. They used only the raw data sampled from the silhouettes associated with the video footage to represent human poses. They also applied the discriminant analysis by spectral regression to project each silhouette in a space of lower dimension. The authors in [114], also proposed descriptors

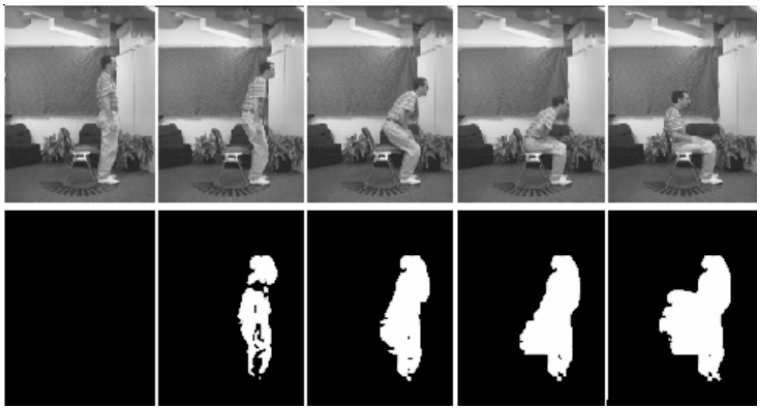


Figure 2.15: Example of someone sitting.

based on silhouettes for their action recognition system. For each action sequence, they first extracted the foreground of the image, then located the silhouette in each image. Then they transformed each silhouette into a time series. Finally, they calculated an aggregated symbolic approximation of the time series, which consists in reducing their size and quantifying them in a set of symbols. The classification of the actions was carried out by the algorithm of random forests. The authors in [8], partitioned Space-Time Volumes (STVs) proposed in [2], into overlapped blocks of a fixed dimension. Afterward, they calculated the 3D-HOG descriptor using Binary data and exploited the 3D vectorial gradients field. Then, each block remains associated with a vector b obtained with the concatenation of SIFT-like [217] histogram of oriented gradients. Figure 2.16 is an overview of their proposed method.

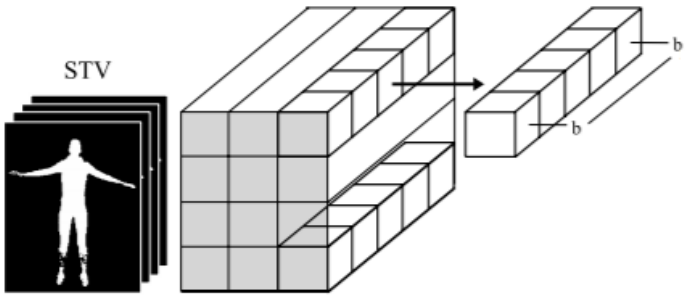


Figure 2.16: STVs and its overlapped blocks partitioning.

- Methods based on skeleton joints: When looking at the representation of a person's movement, identifying the actor can be an important first step. However, the raw shape of the body, is not always discriminating. Certain researchers were inspired by the studies of the researcher in psychology Johanson [113] and focused on the parts of the body of the person for the representation of his movement. This researcher has shown in his work that in order to recognize the actions of the person it suffices just to interpret the trajectories of the part of his/her body. Thus, many researchers have considered this hypothesis in their research, and have represented the human model by a simplified skeleton to identify the positions of its parts. The first solution was proposed in [83] who recovered the pose of the body in 3D from several cameras for tracking and recognizing human movement in 3D space. More recently, several researchers have relied on the skeleton model in their work in order to construct their movement descriptors from information on the skeleton joints [187, 132, 142, 85, 7, 6]. In this context, the authors of [300] presented a new approach for recognizing actions with histograms of joint positions in 3D (HOJ3D) as a compact representation of postures. In this presentation, the 3D space is partitioned into n-bins using a modified spherical coordinate system. The HOJ3D vectors of the calculated training sequences are first reprojected using the linear discriminant analysis method (DAL), then partitioned into k groups with the K-means method to represent the vectors of the movement characteristics. Finally, the Hidden Markov model (HMM) is applied for the classification phase of actions. Jiang et al. [112], proposed a hierarchical model for the recognition of actions which consists in associating each action with a group based on the states of movement of each part of the body. They divided the body into 4 segments (head, abdomen, arms and legs). Then for each group, a model of the k nearest neighbors is trained. It takes as input two types

of characteristics. The first one, presents the motion vector of each joint in a specific time interval. The second one, defines its position relative to a stable joint. Bag of words model is used to represent all the characteristics in order to reduce the size of the descriptor and make the recognition system faster. An adaptive weighting approach is proposed in order to adjust the weight of each word extracted from all the spatio-temporal characteristics and determine the keywords for recognition. The authors in [301], presented a novel descriptor based on a feature-level fusion of spatio-temporal features and skeleton joints. In addition to HOG and 3DHOG, they also added another items to their proposed descriptor using the 3D coordinates of the joints. These items include: a) current posture: pair-wise joint distances in current posture, b) motion: joints difference between current posture and the original (in the first frame), and c) offset: joints differences between current posture and the previous one. In [295], the researchers came up with a new dynamic representation that captures not only the pose of the body in 3D, but also differential properties like the velocity and acceleration of the joints of the human body. The concatenation of this information builds the "Moving Pose" descriptor. An algorithm is adopted to calculate the most discriminating moving images. A voting scheme based on the k nearest neighbor method is used for the classification of the test sequences. Figure 2.17, demonstrates some of the most discriminative frames, automatically discovered by their proposed algorithm, for some actions.

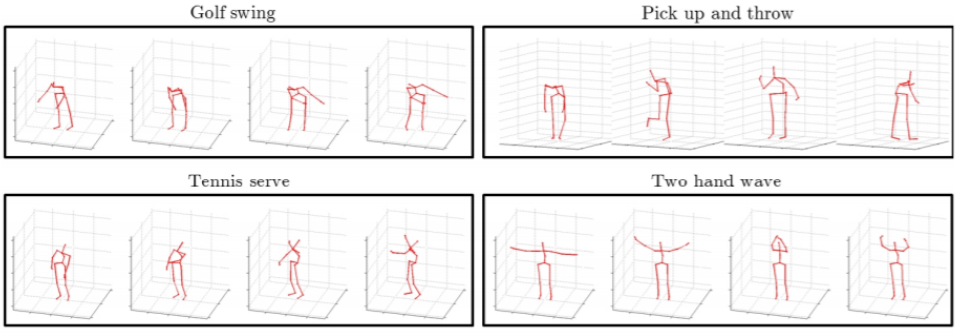


Figure 2.17: Examples of high confidence frames automatically identified from training sequences.

The authors in [106], introduced an approach for the recognition of actions based on the covariance matrix of the positions of the skeleton joints. In order to determine the successive movements, several matrices of covariances were deployed on sub-sequences in a hierarchical manner to define the temporal dependence. The SVM classifier was used for training and classifying actions. In [234], the authors used the 3D point cloud provided by the kinect camera to represent the outer surface of the human body. They presented this surface by describing the relative displacement of neighboring surfaces from a reference point defined in the point cloud. A new cylindrical coordinate system has been defined to make the system invariant to certain possible transformations including translations and rotations. In addition, three diagrams have been proposed to represent the human actions based on the new descriptor, including: a) skeleton-based diagram which defines the difference between two actions by calculating the displacement between two postures, b) scheme of random reference points which consists in sampling an appropriate number of points to cover the body while avoiding possible redundancies by nearby points and c) The spatio-temporal scheme which consists in coding the descriptor in the spatio-temporal domain. The k nearest neighbors method and the one-against-all approach of the support vector machine (SVM) method were used for the

classification of actions. In [287], the authors proposed a new descriptor based on the 3D coordinates of the skeletal joints provided by the Kinect sensor. Their descriptor combines spatial and temporal aspects. In order to explicitly model the displacement, their descriptor called EigenJoints, combines three type of data: the static posture of a pose, the temporal movement of a pose is defined by the difference between the current pose and the previous pose and the offset from an initial pose. After the normalization of these characteristics, the principal component analysis method was applied to have a more compact descriptor. Finally, the non-parametric Bayesian approach was used for the classification of actions. In order to remove the confused images and reduce the cost of calculation in the search for the nearest neighbors, the authors proposed the concept of the accumulated energy of movement which consists in quantifying the distinctive character of each image and therefore selecting the informative images. The authors in [184], developed an open source framework for the recognition of static hand poses as well as their dynamic gestures. The hand segmentation step relies solely on the depth information provided by the kinect sensor and carried out by the Mean Shift segmentation algorithm [56]. As mentioned, their work is divided into two parts. In the first part, recognition of hand poses, training and classification are carried out with support vector machines (SVMs) with radial basis function (RBF) kernels. In the second part, for the recognition of the dynamic gestures of hands, the trajectory of the centroid of the hand was extracted and used as inputs for the Hidden Markov Model in the classification step. Figure 2.18 shows the instance of different hands positions. In [258], the authors proposed a new representation of the skeleton

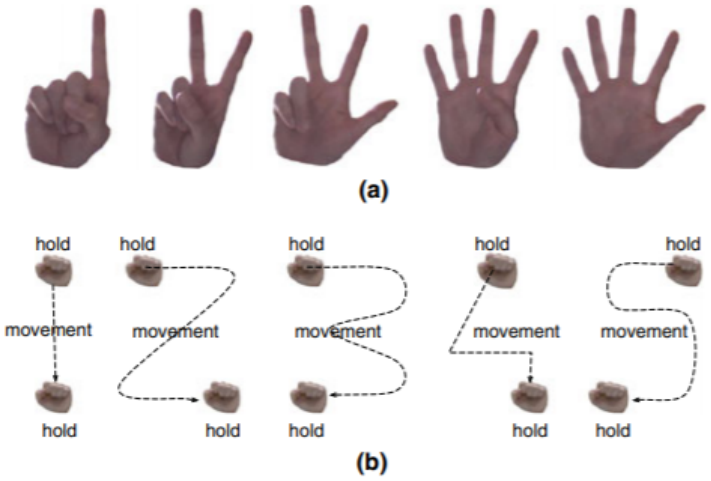


Figure 2.18: Examples of (a) different hand-poses and (b) different hand gestures.

which explicitly models 3D geometric relationships between various parts of the body using rotations and translations in 3D space. The author in [258], proposed a new representation of the skeleton that explicitly models 3D geometric relationships between various parts of the body using rotations and translations in 3D space. Mathematically, the rotations and translations of rigid bodies in 3D space are members of the special Euclidean group SE [171], which is a matrix Lie group. Human actions are subsequently modeled as curves in this Lie group (Fig 2.19). To facilitate the procedure, they associated the action curves of the Lie group with its Lie algebra vector. Finally for the classification of actions, the linear SVM was carried out.

In addition to all these classical methods for recognizing human movements, deep learning and neural

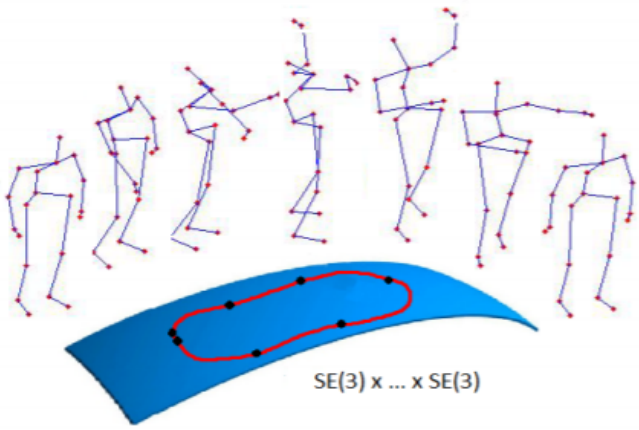


Figure 2.19: Representation of an action (skeletal sequence) as a curve in the Lie group $SE(3) \times \cdots \times SE(3)$.

network methods have also been considered by many researchers in this field. For example in [228], according to the skeletal sequence data, the authors first extracted bone information according to the 2D or 3D joint coordinates. The joints and bones (spatial information) in each frame were then displayed as vertices and edges in a directional circular graph, which is fed into the directed graph neural network (DGNN) to extract features for gesture detection. They implemented their proposed algorithm on two large public data-sets, NTURGB+D [151] and Skeleton-Kinetics [117], and achieved significant results. In [229], the authors considered three main factors that contribute to the complexity of the pattern in a movement, including spatial dependence between joints, body temporal dependence, and changes in performance such as velocity and accelerations. Their proposed solution offers a combination of graph diagram and Long-Short term memory (LSTM) for space-time dynamics modeling. The whole problem then extends to be a probabilistic model following Bayesian framework with a novel adversarial prior. In order to improve robustness and increase detection accuracy, a Bayesian inference problem has been designed for classification phase. In [144], the authors calculated the latent and subsurface dependencies of all connections using actional links (A-links) to discover the deeper dependencies of human skeletal joints. They also expand skeletal charts to show higher-order relationships as structural links (S-links). By the use of their proposed actional structural graph convolution network (AS-GCN), they succeed to have a significant result on NTURGB+D data-set. In [188], the researchers codify the spatio temporal pattern of a gesture into color images. Afterward, by the use of a DenseNet algorithm, in which each layer is connected to all the others within a dense block and all layers can access feature maps from their preceding layers, succeed to train and classify gestures via proposed motion representation. Their proposed algorithm was implemented in two public data sets: MSR Action3D [145], NTURGB+D and CEMEST [188] data-set. Their results outperform most of the studies in the literature. In [284], the authors tried to analyze skeleton sequence data via a Double-feature Double-motion Network (DD-Net) for skeleton-based action recognition. Their proposed descriptor consisted of the distances between the joints and their velocity. The factors in this descriptor are invariant to the point of view of the cameras and the location of the participants. They used SHREC [38] and JHMDB [111] data-set, in order to evaluate their proposed method. Since generally, the movements of the body might have a different attribute due to the orientation misalignment, the authors in In [138], transformed the original raw data which is the 3D

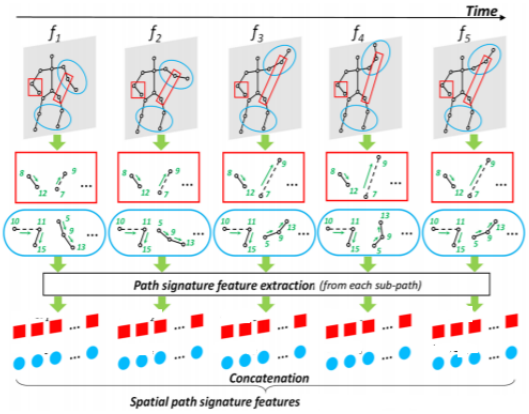


Figure 2.20: The illustration of spatial feature .

coordinate of the joints into a human cognitive coordinate system. Afterward, by the use of the multi-term temporal sliding LSTM networks, they fed different type of dependencies including spatio and temporal data into network and at the end by implementing their proposed method on MSR Action3D, UTKinect-Action [127], NTURGB+D [224], Northwestern-UCLA [263], and UWA3DII [197] data-sets, they evaluate their proposed algorithm. In [29], the authors demonstrate the 3D coordinates of body skeleton as a point in the Kendall's [119] shape space so that a gesture can be shown as a trajectory in 3D space. With this representation, they had the ability to focus on the important invariance properties by analyzing its shape. They use two pipeline in order to evaluate their proposed gesture representation on Florence3D-Action [220], UT-Kinect Action and MSR Action 3D. They first try the Dynamic Time Warping (DTW) and Support Vector Machine (SVM) and then the LSTM method in order to train and evaluate their data. In [285], the authors, considered the 3D coordinates of all the joints, as well as the pose disintegration with $m = 2$ and $m = 3$, which means joint pairs and joint triples are used as illustrated in Fig 2.20. Also, to represent temporal feature, they calculated the path traveled by all joints during a gesture. By the concatenation all spatial temporal features, they prepared the input of a linear single-hidden-layer fully connected network in order to classify the gestures. Four data sets: HMDB [111], SBU [294], Berkeley MHAD [177], and NTURGB+D [224] were used to evaluate their proposed pipeline.

Laban Movement Analysis (LMA) [134], is another method which uses temporal and spatial information to formally describe human movement. This algorithm observes the gestures based on four aspect of the movement, namely body, effort, space and shape. This method, which can accurately describe the movements due to considering all aspects of a gesture, has been considered by many researchers. This method, which can accurately describe the movements due to considering all aspects of a gesture, has been considered by many researchers. For example, in [268], this method is used to construct a descriptor for dance analysis. They used spatial orientation, limb structure and force effect to calculate the components of LMA. In training and classification phase, they proposed the CNN-LSTM hybrid deep learning model, which is a combination of two network structures of LSTM and CNN, and verifies the effectiveness of the method through contrast experiments. Their results show that the CNN-LSTM model has the highest accuracy rate. In [7], the authors used LMA to classify human emotions based on body movements. They used Random Decision Forest to classify the emotions.

The authors in [175], modeled a three-dimensional block whose dimensions are, respectively, the number of skeleton joints, the number of consecutive frame, and the three spatial coordinates (x, y, z) of the joints. They combined the Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) recurrent network as a machine learning method to recognize human action and hand gesture. The researchers in [267], reshaped the 3D spatio-temporal data into three texture 2D images through color encoding, Joint Trajectory Maps (JTMs), and implemented Convolutional Neural Network to train the discriminative features for classifying human gestures. In [244], based on the 3D coordinates of the joints, a graph-based structure is proposed for gesture recognition. In this work, the joints and their dependency were considered as a graph. Vertices of the graph contain the 3D coordinates of the body joints, while the adjacency matrix captures their relationship. In [49], the authors used the joint angles and orientations of the most informative body parts to define their descriptor (Fig 2.21). Because they evaluated the proposed descriptor in small-sized data-sets, they used Support Vector Machine (SVM) for training and classification.

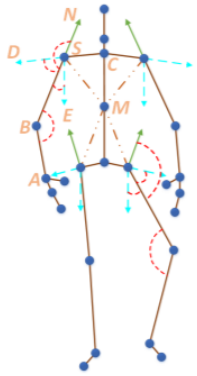


Figure 2.21: Illustration of the joint angles of the skeleton.

So as we can see different descriptors have been proposed for the representation and classification of movements. In addition to this, an increasing number of RGB-D databases have been built for use in the evaluation of these algorithms. The use of publicly accessible databases not only saves time and resources for researchers, but also allows for a fair comparison of different systems. Each database is based on specific criteria, such as the type of gesture (metaphorical, iconic, etc.), the complexity and similarity of the gestures, the change of point of view, etc. In the next part we will present some public databases used in our thesis.

2.4 Public human action data sets

In our work, we evaluated our system with 5 public databases to ensure the robustness of our approach, which are MSR Action 3D, UTKinect, Florence 3D, SYSU 3D HOI and NTU RGB+D 120.

2.4.1 MSR Action 3D

The MSR Action 3D database was introduced in [145] and contains 20 different actions (Upward sign, horizontal sign, hammer strike, wave with one hand, punch forward, throw away, draw an X, draw a check mark, draw a circle, clap, wave with both hands, box to the side, bend over, kick forward, kick side walk, jog, tennis swing,

tennis serve, golf swing, and pick up and throw), performed by 10 different people. Each action is repeated 2 or 3 times, in total there are 567 sequences. The 3D positions of 20 joints were captured using a depth sensor similar to the Kinect camera with a resolution of 640×480 pixels (15 frames per second). All the videos are recorded from a fixed point of view and all the participants were in front of the camera while carrying out the actions. The background has been removed from the database during pre-processing. Data is provided as segmented samples. The database is divided into 3 groups, AS1, AS2 and AS3, each consisting of 8 actions as shown in Table 2.2. The AS1 and AS2 subsets were intended to group similar actions, while AS3 was intended to group all complex actions. Figure 2.22, presents some instance of the actions from the MSR Action 3D database. And Figure 2.23 illustrates the 3D silhouette movements for the draw tick mark and tennis serve gestures.

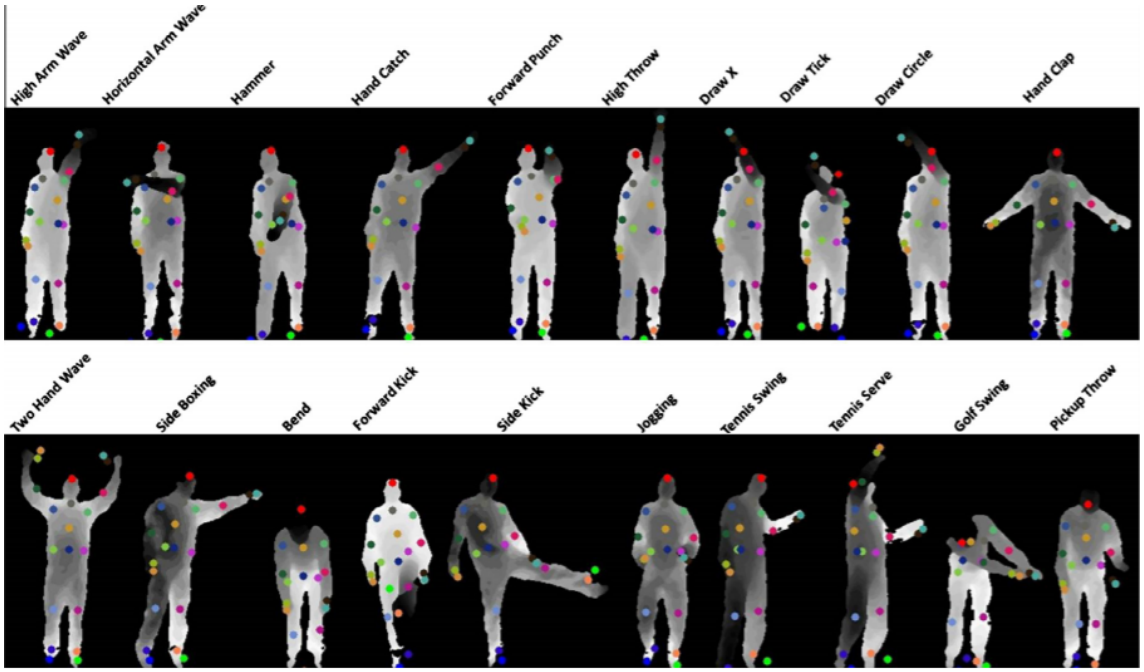


Figure 2.22: Examples of depth maps and skeleton joints associated with each frame of twenty actions in the MSR Action3D dataset [286].

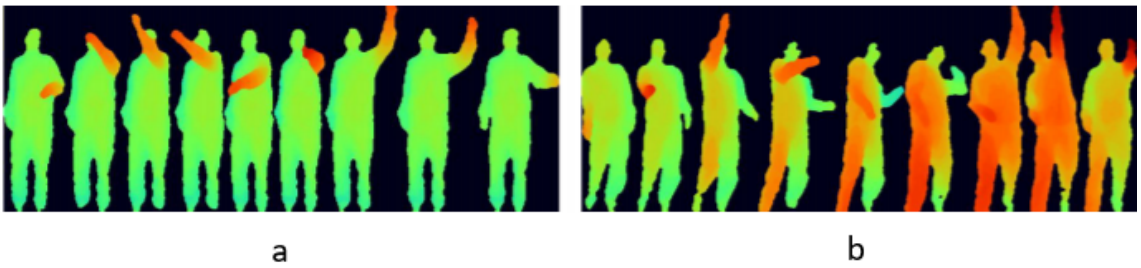


Figure 2.23: Illustration of the 3D silhouette movements for the draw tick mark and tennis serve gestures [145].

2.4.2 FLORENCE 3D ACTIONS DATASET

The Florence 3D Actions data set [19] collected at the University of Florence during 2012, has been captured using a Kinect camera. It includes 9 activities: wave, drink from a bottle, answer phone,clap, tight lace, sit down, stand up, read watch, bow. During acquisition, 10 subjects were asked to perform the above actions for

Table 2.2: The classes of gestures of the MSR Action 3D database and their number of repetitions (NR)

AS1	NR	AS1	NR	AS1	NR
Horizontal arm wave	27	High arm wave	27	High throw	26
Hammer	27	Hand catch	26	Forward kick	30
Forward punch	26	Draw X	28	Side kick	30
High throw	26	Draw tick	30	jogging	30
Hand clap	30	Draw circle	30	Tennis swing	30
Bend two	30	Hand wave	30	Tennis serve	30
Tennis serve	30	Forward kick	30	Golf swing	30
Pick up and throw	30	side boxing	30	Pick up and throw	30

2/3 times. This resulted in a total of 215 activity samples. Figure 2.24 shows the movements performed in this data set.

2.4.3 SYSU 3D HUMAN-OBJECT INTERACTION data set [102]

For constructing this data set, 40 subjects were asked to perform 12 different activities such as drinking, pouring, calling phone, playing phone, wearing backpacks, packing backpacks, sitting chair, moving chair, taking out wallet, taking from wallet, mopping, and sweeping. For each activity, each participants manipulate one of the six different objects: phone, chair, bag, wallet, mop and besom. Therefore, there are totally 480 video clips collected in this set. For each video clip, the corresponding RGB frames, depth sequence and skeleton data are captured by a Kinect camera. Sample activities are shown in Figure 2.25. Compared to existing data sets, this data set presents new challenges:

- The involved motions and the manipulated objects' appearance are highly similar between some activities.
- The number of participants is at least four times larger than that of existing ones.

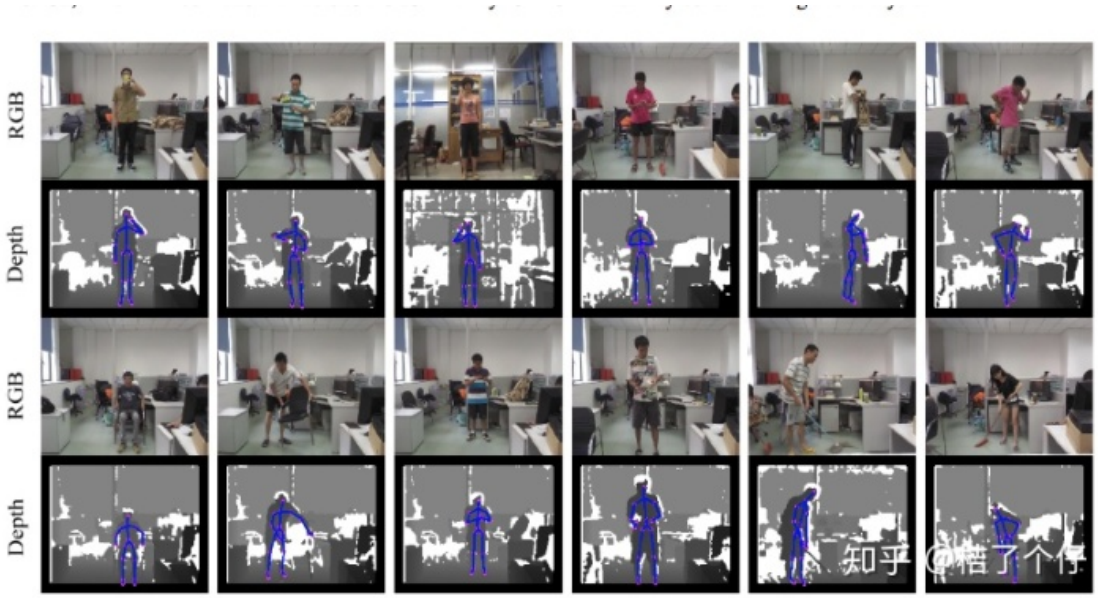


Figure 2.25: Snapshots of activities in SYSU 3D HOI set, one sample per class. The rows headed with RGB show the samples in RGB channel and the rows underneath headed with Depth show the corresponding depth channel superimposed with skeleton data. Best viewed in color.

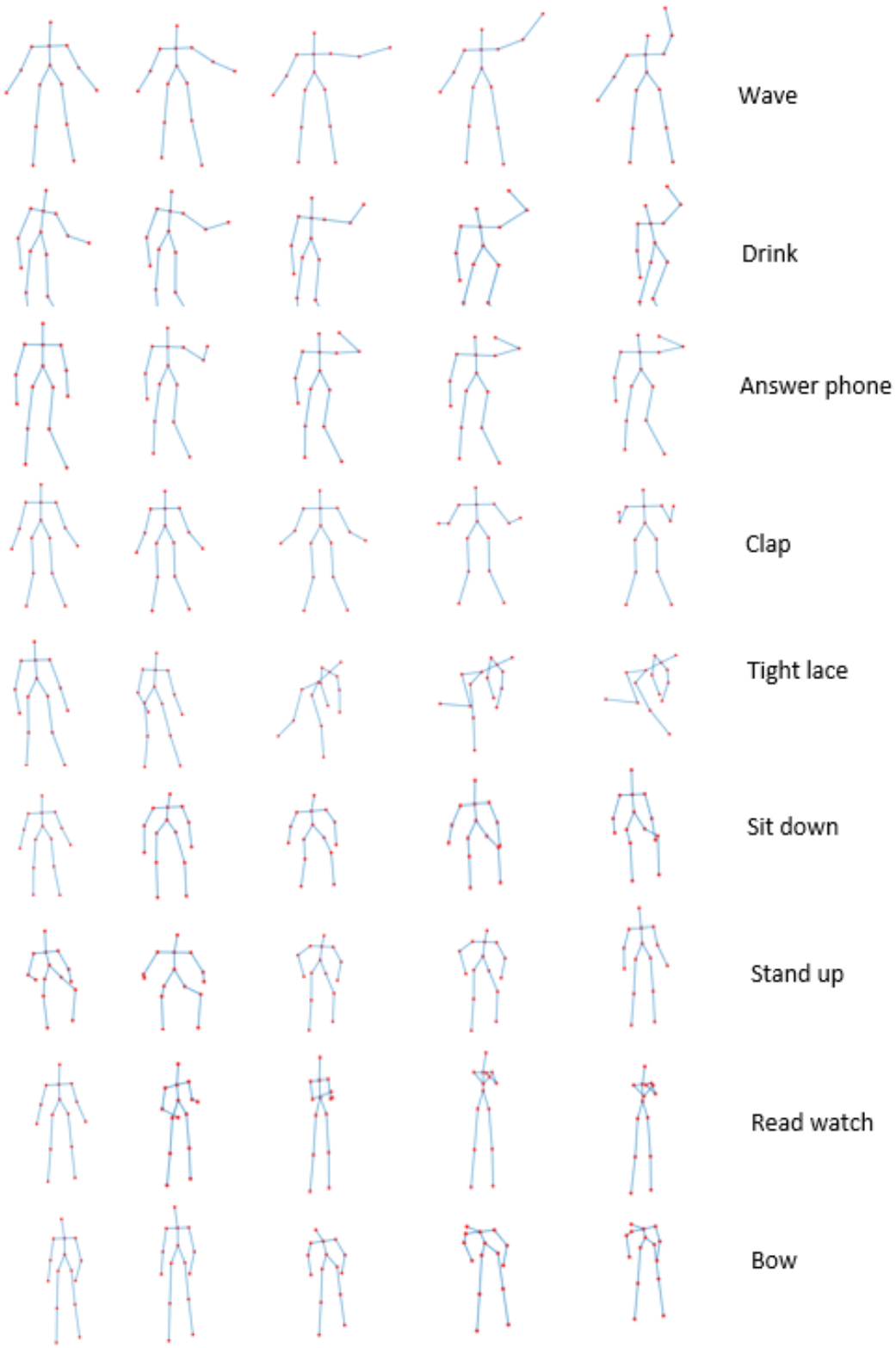


Figure 2.24: Skeletal representation of all movements performed in FLORENCE 3D ACTIONS DATASET.

2.4.4 UTKinect-Action3D Dataset [279]

This data set was collected as part of research work on action recognition from depth sequences. The videos was captured using a single stationary Kinect with Kinect for Windows SDK Beta Version. There are 10 action types: walk, sit down, stand up, pick up, carry, throw, push, pull, wave hands, clap hands. There are 10 subjects, each subject performs each actions twice. Three channels were recorded: RGB, depth and skeleton joint locations. The three channels are synchronized. The frame rate is 30f/s. Note we only recorded the frames when the skeleton was tracked, the frame number of the files has jumps. The final frame rate is about 15f/sec. Sample activities are shown in Figure 2.26.

Figure 2.26: Snapshots of some activities in UTKinect-Action3D Dataset.

2.4.5 NTU RGB+D 120 [151]

Microsoft Kinect sensors is used to collect this dataset. Four major data modalities are acquired by this sensor, namely, the depth maps, the 3D joint information, the RGB frames, and the infrared (IR) sequences. The depth maps are sequences of two dimensional depth values in millimeters. To maintain all the information, lossless compression for each individual frame, is applied. The resolution of each depth frame is 512×424 pixels. The joint information consists of 3-dimensional locations of 25 major body joints for each detected and tracked human body in the scene. The corresponding pixels on RGB frames and depth maps are also provided for each body joint. The configuration of these joints is illustrated in Fig 2.27. The RGB videos are recorded in the provided resolution of 1920×1080 pixels. The infrared sequences are also collected and stored frame by frame at the resolution of 512×424 pixels. It consists of 120 action categories in total, which are divided into three major groups, including 82 daily actions (eating, writing, sitting down, moving objects, etc), 12 health-related actions (blowing nose, vomiting, staggering, falling down, etc), and 26 mutual actions (handshaking, pushing, hitting, hugging, etc).

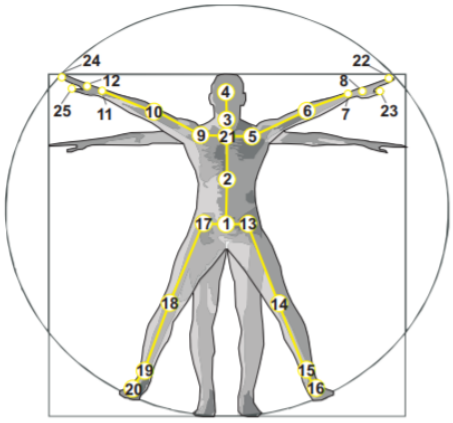


Figure 2.27: Illustration of the configuration of 25 body joints in NTU RGB+D 120. The labels of these joints are: (1) base of spine, (2) middle of spine, (3) neck, (4) head, (5) left shoulder, (6) left elbow, (7) left wrist, (8) left hand, (9) right shoulder, (10) right elbow, (11) right wrist, (12) right hand, (13) left hip, (14) left knee, (15) left ankle, (16) left foot, (17) right hip, (18) right knee, (19) right ankle, (20) right foot, (21) spine, (22) tip of left hand, (23) left thumb, (24) tip of right hand, (25) right thumb.

2.5 Emotional language

2.5.1 Definition of an emotion

The word "emotion" comes from the French word "émouvoir". It is based on the Latin emovere, "e" means "out of" and "movere" means "movement". The related term "motivation" is also derived from the word movere. In general, we can say that an emotion is a psychological and physical reaction to a situation. It first has an internal manifestation and generates an external reaction. It is generated by the confrontation with a situation as well as by the interpretation of reality. However, emotion always remains specific to each individual [189]. Till now, there is no agreement on what an emotion is [215], [79]. In a recent survey, internationally renowned experts in emotion research were asked to define emotion. As expected, there was indeed no consensus [109]. Consequently, several definitions and roles have been given to emotion [165], [181]. In 1879, Charles Darwin, founder of the theory of evolution, defined it as an innate, universal and communicative quality, linked to the past of the evolution of our species. However, there is still a consensus regarding the view that emotions have more than one psychological or behavioral manifestation: in addition to subjective feelings, they also contain tendencies to action, physiological arousal, cognitive assessment and expressive behavior [235]. Emotions are generally viewed as one of many different effective phenomena intrinsic to the human experience such as mood, interpersonal position, attitude and personality traits [215]. All of these affective phenomena have the power to cause changes in human physiology. They can be distinguished according to a number of dimensions including intensity, duration and degree of coordination between different modalities. The most well-known model of classification of emotions is that of James Russell [206], called the Circumplex model of affect (see Figure 2.28).

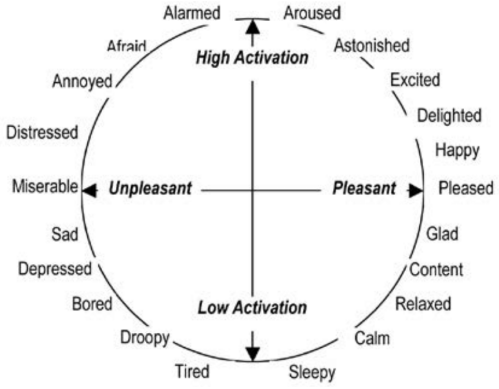


Figure 2.28: Circumplex model of affect.

This model makes it possible to describe the emotion in a two-dimensional affective space represented by two axes: a horizontal axis corresponds to the dimension of Valence which makes it possible to distinguish between positive and negative emotion and a vertical axis corresponds to the Arousal dimension which defines the intensity emotion and differentiates an active emotion from a passive emotion. Emotion can be expressed by different modalities, the most processed are: speech, facial expressions and the body.

2.5.1.1 Emotions expressed through words

Much research has been devoted to the study of automatic recognition of emotions through the analysis of human speech [30, 58, 16, 259, 214, 232]. Some of this research has been applied to call centers, multi-agent systems or other areas such as [186, 289, 154, 255, 154]. Speech recognition requires the extraction of relevant features. The most commonly used characteristics are of an acoustic or prosodic nature (vocal energy, speech rate, frequency, duration, etc.) and therefore capture the quality of speech to identify the associated emotion. However, most of this research is only for speaker dependent recognition. Recognition of speaker independent emotion is a difficult task. In a survey conducted to measure human performance on emotion recognition, only 60% of people can correctly determine the emotions expressed by unknown people [219]. Some authors have shown that this modality requires adaptation of the speaker. For example, the authors in [261], have shown that adding a gender detection step in the emotion recognition system leads to better performance. Other authors have stressed the importance of speaker normalization for the recognition of emotions [223, 260]. Other constraints can be introduced into a system for recognizing emotions via speech, such as the case of communication with deaf people, of a long-distance conversation or in a noisy environment. In such situations, the system will not always be able to transmit emotions through the vocal channel.

2.5.1.2 Emotions expressed by the face

For non-verbal communication, facial expressions have been considered as the main modality used to convey emotions [206]. The facial expressions of an emotion are a consequence of the movement of the muscles under the skin of our face [68]. The movement of these muscles causes the deformation of the facial skin in a way that an external observer can use it to interpret the associated emotion. Each muscle used to create these facial constructions is called a unit of action (AU). The authors in [70],identified the AUs responsible for generating the emotions most often observed in most cultures: anger, sadness, fear, surprise, happiness and disgust. For the

coding of these facial expressions, they developed the FACS (facial action coding system), a system for coding visible facial movements, to encode facial movements and provide an indication of the degree and intensity of activation of muscles. This system has been widely exploited by several researchers for the recognition of emotions through facial expressions. In 1990, the same authors [62] proposed Duchenne's smile (D) as a spontaneous and authentic expression of positive emotions, such as happiness, pleasure, etc. They classified the smile (D) as a combination of two muscles: the zygomatic major muscle (AU 12) which pulls the corners of the lips upwards thus producing a smiling mouth and the orbicularis oculi (AU 42), a muscle located around the eyes, to lift the cheeks, narrow the eye opening and form wrinkles around the eye socket. However, during social interactions, some researchers have found that emotional perception through the face can be influenced by certain factors, for example the long distance where facial expressions become not too clear, or also the age factor. Finally, there is also another important factor which is the context. For some authors, the perception of emotion through facial expressions is biased in the direction of the expressions bodily [17, 63, 63, 212, 201, 268]. Another modality has appeared in this area which is the movement of the body. The role of gestures in the perception of emotion has become of great importance.

2.5.1.3 Emotions expressed by the body

Psychological researchers were the first to focus on bodily expressions according to posture and body movement. The latter support the idea that bodily expressions "speak" more than facial expressions [164]. We are talking here about expressive gestures, which are gestures made in an emotional state. This idea has attracted the attention of many researchers in recent years [77, 76]. Expressive gestures are applied in several fields, including dance, music, animation, etc. The researchers in [47], developed software called Eyesweb (See Figure 2.29) in the InfoMus laboratory at the University of Genoa to facilitate real-time analysis of expressive dance gestures.

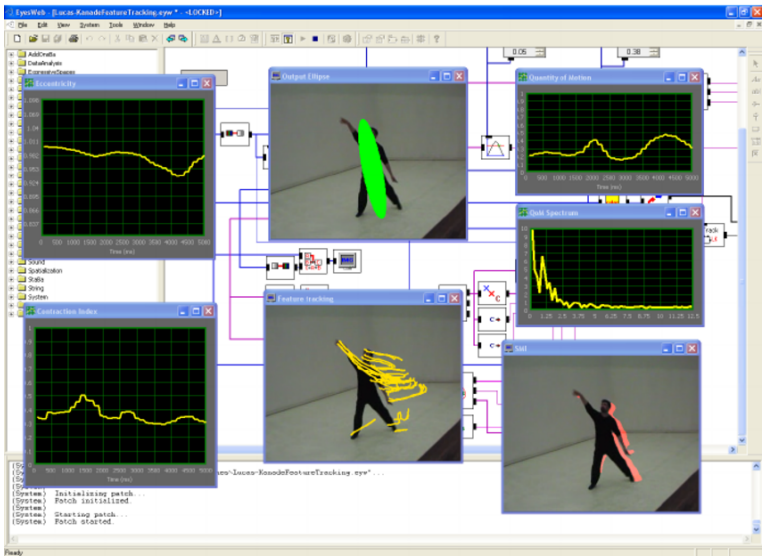


Figure 2.29: An EyesWeb application extracting motion cues (QoM, CI, Kinematics cues) [46]

They identified movement cues thought to be important for emotion recognition and investigated how these cues can be tracked by automatic recognition techniques. They adopted a layered approach [140] to model movements from low-level physical measurements (eg, position, speed, acceleration of body parts) to descriptors of overall movement characteristics. (eg fluidity of movement, openness, impulsiveness). They finally showed

how these clues help transfer four emotions (anger, fear, grief and joy) into a choreography. On the basis of these movement indices, the authors defined an automatic classifier capable of distinguishing the four states. Expressive gestures are also used in the animation of conversational agents [94]. The idea here is to modify the gestures by adding expressiveness to them in order to increase the credibility of the Agents and the naturalness of their behavior. They are found in the field of style transfer which consists of transforming a movement sequence into a new movement style while keeping its original content. In this context, the authors of [101], used the temporal distortion method to transpose the input movement sequence into the same expressive movement sequence. The authors in [280], designed an animation system to add different styles to existing animation. They constructed a series of local blends of autoregressive models to represent the complex relationships between styles and automatically transform an unlabeled heterogeneous sequence of movements into different styles. The authors in [75], constructed a new data set of expressive daily activities, named Emylia. Their data set contained many widely used emotions in body movements. It includes synchronized audio-visual and motion capture recording. In this data set, video files including emotional states of the face and body are viewed from different perspectives using the camera, and motion recording files include three-dimensional data of the entire body. In order to evalute this data set, its authors used statistical method based on human approach. The results have shown that the emotions Anger, Sadness, Neutral, Panic Fear and Joy were significantly better perceived in the videos showing the expression of the same emotions than from the videos showing the expression of other emotions. Shame perception received also significant mean rating, but the expression of Shame was confused with the expression of Sadness. Confusion was also found among the expression of Pride and Joy as well as the expression of Anxiety and Shame. The authors in [293], proposed an approach for the transfer style based on spectral analysis, which manages heterogeneous movement sequences and also transfers the style between independent actions. The characterization of movements in general revealed two levels of descriptors: low-level and high-level descriptors. These descriptors are based on information related to the joints, such as 3D position, speed, acceleration, angle of rotation, etc. High-level descriptors depend on the context of the action and require a well-defined formalism that allows the movement to be described while considering its context. In this context, a model appeared called the LMA model (Laban Movement Analysis) developed by Laban [134]. This is an approach initially used in the study of dance movement with an approach centered on the quality of the movement. This formalism thus makes it possible to know how to describe a movement in a complete way and with the minimum number of characteristics sufficient.

2.6 The model of Laban Movement Analysis

In order to achieve a satisfactory simulation for the complex language of the human body, as simple as possible but much necessary complex description of human movement is needed and LMA fulfills these demands. This method consists of describing, visualizing, interpreting and documenting human movement. It uses a multi-layered description of movement, focusing on its four components (Body, Space, Shape and Effort). The Body component is used to describe human movement. Space is used to describe the trajectory made by the parts of the body during the execution of a movement. Shape deals with the change in body shape during a movement based on three factors: shaping, shape flow and directional movement. Finally, the Effort component describes

Table 2.3: The four components of LMA with their factors.

LMA	Body			
	Space			
	Shape			
	Shape	Shape flow		Directional movement
	Effort			
	Space	Time	Weight	Flow

the expressiveness and quality of the movement following four factors, straightness (space), rapidity (time), force (weight) and fluidity (flow) of movement. Table 2.3 illustrates the different factors of the LMA method.

- The Body component determines what is in motion, what parts are connected, which parts are influenced by others, and the order or sequencing of movements. By focusing on the way the body is used, it is possible to differentiate between gestures that involve isolated parts of the body, postures that are supported by the body and whole body actions such as actions: jumping, running, stretch or twist. With regard to motion sequencing, the observer can distinguish between simultaneous (two or more parts at a time), successive (adjacent body parts one after the other), sequential (non-adjacent parts) sequencing one after the other) or unitary (whole body).

- The Space component describes in which space the movement takes place. Laban defines the kine-sphere as a personal imaginary space placed around the person and accessible directly by his limbs to the tips of the fingers and feet stretched in all directions.

- The Shape component analyzes how the body changes shape during movement. It was developed in the years 1950-1960 by Rudolf Laban and Warren Agneau to observe and work on the structural transformations of the human body in a three-dimensional space, linked to oneself and to the environment. It addresses the following three questions:

  - What forms does the body make?

  - The shape change in relation to itself or in relation to the environment?

  - How does the shape change?

  To answer these three questions, the category of Shape implies three distinct qualities of change in the form of movement: shape flow, directional movement, and shaping. Shape flow characterizes the change in body shape. It can be described by an attitude of internal surrender which is intimately linked to breathing (Fill / Empty). Directional movement defines the path of movement in space, which can be rectilinear or curvilinear. For example, pointing or pushing an object are linear movements. In contrast, swinging a tennis racket or painting a fence present bent movements. Shaping represents the relationship between the moving body and 3D space. Shape changes in movement can be described in terms of three dimensions: horizontal, vertical and sagittal. Each of these dimensions is in fact associated with one of the three main factors (width, length and depth) as well as one of the three planes (horizontal, vertical and sagittal) linked to the human body. Shape changes in the horizontal dimension occur mainly in

the lateral directions. The changes in the vertical dimension are mainly manifested in the up and down directions. Finally, the changes in the sagittal dimension are more evident in the depth of the body or the front-to-back direction.

- The Effort component describes the expressiveness of the movement according to its four factors:
  - Space: describes the directivity of movement between two qualities (Direct and Indirect).
  - Time: describes the speed of movement between two qualities (Sudden and Sustained).
  - Weight: describes the force of movement between two qualities (Strong and Light).
  - Flow: describes the fluidity of movement between two qualities (Bound and Free).

The Effort-Forme subdomain has received considerable interest in its way of describing the qualities of movement. This duet makes it possible to describe the quality and the rhythm of the movement as well as the expressiveness of the gesture. The LMA method has been used in the literature for several purposes, including:

- Gestural animation and synthesis of expressive gestures: the authors of [53], have developed a system for generating expressive gestures called the EMOTE model (Expressive MOTion Engine). This system consists in modifying a predefined movement in order to produce expressive gestures for the virtual agents. They used the Effort-Shape components of LMA to produce expressive movements in the upper body (torso and arms). The Shape parameters are applied to the arms and torso, while only the arms are concerned with the Effort qualities. Likewise, the authors of [299], used the results of the EMOTE model in order to perform gestural animation. They used a hidden single-layer neural network with the gradient back-propagation approach to detect movement characteristics from the gestures performed and estimate the relationships between these characteristics and the qualities of Effort. The combination of their system with the EMOTE model makes it possible to automate the processes of human observation and analysis and to produce natural gestures for communication agents from 3D motion capture. The researchers in [69], have improved the EMOTE model proposed by [53] for the purpose of studying human personality through the movement of his body. In order to characterize the dynamic aspects of the movement, the authors derived physical measures of the factors of the force component. The OCEAN model has been adopted to define the 5 main personality traits (Openness, Consciousness, Extra version, Pleasure and Neurosis). Figure 2.30 shows which Shape Quality affects which effectors. An association is established between Stress parameters and OCEAN factors to find the relationship between personality and the Stress component of LMA. This relation was used for a generalization of the representation of the personality through the movements. They considered applying this relation to produce a stylized variation of the movement of a virtual agent by adjusting the parameters of Effort.

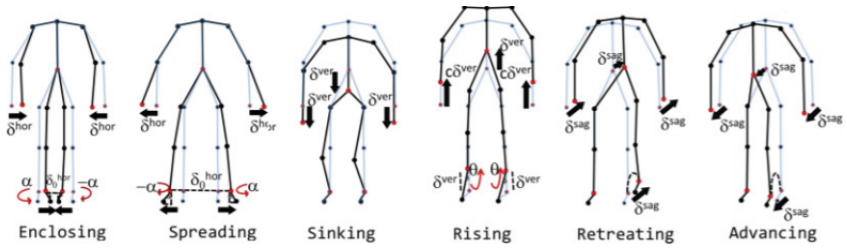


Figure 2.31: Sample frames from motion captured folk dances

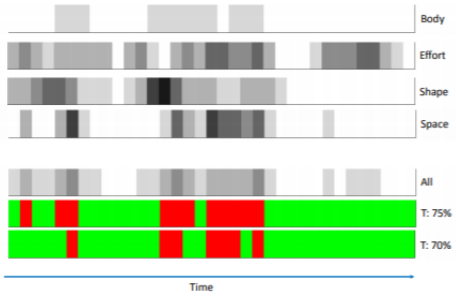


Figure 2.32: The correlation between the movements of the teacher and student; the first four bars show the correlation for each LMA component separately, while the next shows the overall correlation taking into consideration all the LMA components. The correlation is presented in grayscale, where white means high correlation and black means no correlation. The last two bars show the decision whether the movements under investigation are similar or not, when the passdecision threshold is set at 75% and 70% respectively. Green means “pass", while red mean “fail".

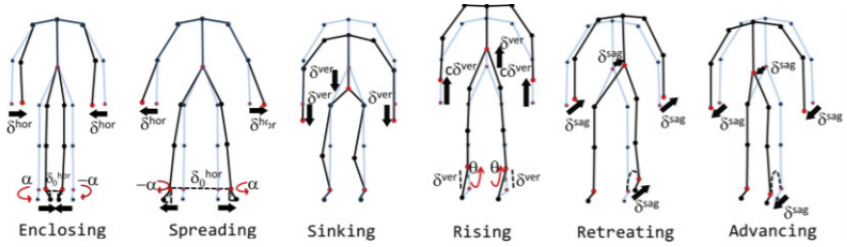


Figure 2.30: Shape Qualities and their application on effectors of the inverse kinematics solver. Red dots are the effectors that are explicitly updated by the Shape changes. Black arrows and red curves show translational and rotational changes, respectively

– Motion indexing and retrieval: This approach consists of motion retrieval based on its content from the dataset. The LMA method is used to encode the characteristics of the movement. These characteristics provide a representative search space for indexing movements. In this context, Kapadia et al. [116], proposed a compact representation that sufficiently captures the characteristics of human movement and provides an efficient means for indexing movements regardless of the size of the data set. Their characteristics were inspired by three factors of LMA (Body, Effort and Shape). These characteristics are then combined to find complex movements in large movement data sets. The authors in [11], used the LMA method to extract indicative characteristics for dance movements 2.31. They studied the correlation between different emotions based on the characteristics of movement 2.32 . The same authors [10] used the same descriptor and proposed a search algorithm which consists in measuring the correlation between the different movements by relying on the characteristics of this descriptor. This makes it possible to find the potential similarities between the different dance clips and thus recover the movements with similar qualities.

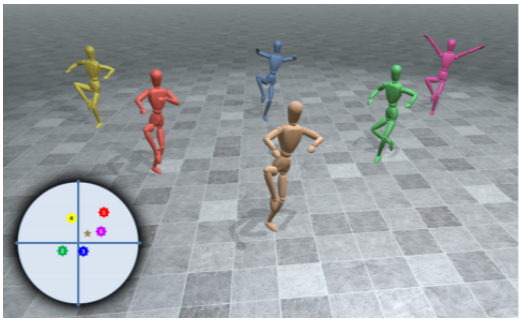


Figure 2.33: : Six characters dancing with different emotions. The emotion coordinates are visualized in the lower-left inset with corresponding colors. The brown character shows the original motion. The two characters in the middle dance more happily (red) or more sadly (green). The three dances at the back are edited towards afraid (yellow), tired (blue), and pleased (purple), respectively.

– Style Transfer: Style transfer methods examine the issue of transferring the style of a movement from one person to another. The authors in [249], resulted in a non-linear mapping between animation parameters and movement styles in perceptual space. This mapping can then be used to synthesize variations stylistic from artificially generated examples using Effort factors of LMA. The authors in [14], used the LMA technique to synthesize the movements humans from existing motion capture data. They extracted quantitative and qualitative characteristics of movement based on the components of LMA. They applied the RBF regression model to relate motion characteristics to their emotion coordinates on Russell's circumplex model 3.2. This allows the movements to be stylized with emotions by modifying the selected characteristics. The same authors [15] have extended LMA-based framework by adding other features that take into consideration the modes of interaction with itself, others and the environment, aimed at improving stylistic consistency by relation to the spatial component which has not been fully studied found in [14]. They first extracted the movement characteristics inspired by the components of LMA and found their stylistic correlations. They built a motion graph based on the correlations between postures in order to find potential transitions between videos. These correlations are used to prune motion graph transitions that are not stylistically consistent, leading to an LMA-based motion graph. This new motion graph can be used to synthesize plausible dance animations.

– Quantification of the expressive content of gestures in relation to emotion: the authors in [211], adopted the LMA method for motion analysis hands and arms. They quantified the factors of Effort (space, time, weight and ux) as well as the factor of directional movement of Shape based on measured movement characteristics, such as speed, acceleration, etc. In their database, six hand and arm movements were performed by a professional actor to convey six emotions (anger, happiness, sadness, fear, disgust and surprise). Subsequently, their database was annotated by a Certified Motion Analyst (CMA) to study the statistical correlation between CMA annotations and quantified LMA factors. The authors in [251], used the LMA method to classify the conductors' gestures and also analyze the expressive content of their gestures. In their database, the authors recorded 8 different sessions. They then segmented each session manually in order to create samples of the gestures. These gestures are then annotated by experts by selecting emotions from the categories

offered in their data set. The researchers in [13], used the LMA method to encode the physical and also stylistic characteristics of the movement. They have developed a folk dance learning platform to help beginners learn this dance by following a 3D avatar. Professional dancers have been invited to build their database. Each user therefore imitates the dance of the avatar. Its movement is analyzed and compared to that of the model. This comparison is based on the qualities extracted from the user and the avatar inspired by the LMA method. Finally, an evaluation of the user's performance is deduced.

2.7 Our approach

As we can see, most of the work on expressive gestures has often turned to dance or music. These dance sequences tend to be continuous, without the artificial restriction of having to start and end with a neutral pose. For this, most of the work has focused on the analysis of expressive gestures without going through training and classification. Another limitation in this field is the creation of human gesture data sets, which basically require training courses as well as specialists in this field. As part of a Human-Robot Interaction (HRI), we found a few papers that involved the LMA model for different purposes. Some have used this model to characterize the trajectories of robots, [130], which generated expressive movement trajectories of the robot based on the Effort component of LMA. They extracted 3 measurements (x, y position and θ orientation) and created a motion descriptor vector based on these characteristics to quantify each factor of the Effort component. Finally, they studied the relationship between 6 emotions and the parameters of Effort. The same authors in [131], quantified the Space factor of the Effort component to study the interpretation of people on the attitudes of the robot in relation to its point of arrival (hesitation, direct, lost track of goal) through the characteristics of the trajectory of its movement. The authors of [226], were based on the component Effort of LMA to characterize the trajectory of the movement of a flying robot. They recruited a trained artist in Laban to create a set of movements for each combination of Effort parameters (space, time, weight and flow). The 4 parameters with their two extreme qualities then give 16 combinations. Then, they adopted the Circumplex model to associate the affect perceived from robotic movements on the two dimensions: valence and excitation. Another type of research in the HRI field based on the LMA model consists in generating expressive humanoid robot gestures by varying the parameters of the LMA factors. Another example is [123] who proposed a computational model of the factors of weight and time and applied it to robotic platforms to develop a method of diversification of gestural movements of the Darwin-OP robot. Also, the authors in [159, 160], used the qualities of Laban to give expressive gestures (pleasure, anger, sadness and relaxation) to the humanoid robot KHR-2HV. Their method is to add a target emotion to arbitrary body movements of a human-shaped robot while modifying Laban's parameters. Other researchers have considered that the interpretation and recognition of a person's movement by a robot makes human-robot interaction more natural. They therefore applied the LMA model for the characterization of user gestures, such as [126], which represented the emotional movements of the human body with the three factors of the Effort component. They quantified the three factors (space, weight and time) to characterize two emotional movements (rejoices and complains). Finally, we cite the work of authors of [155, 21] who considered that the perception and interpretation of non-verbal behavior by a robot

is important for a natural Human-Robot interaction. Two characteristics (acceleration and frequency) are extracted to quantify the Effort component in order to characterize the gesture "making a sign with a hand" performed with 4 emotions (joy, anger, sadness and politeness). Their goal is to develop an NAO robot capable of recognizing and imitating human movements and thus ensure a natural interaction with autistic children within the framework of a collective game. The same authors in [124], recently developed a framework for robot-assisted training of children with autism spectrum disorders (ASD). Their framework is created using the Choregraphe software developed by the Aldebaran company, which makes it possible to create the behavior of the NAO robot with a graphic language. The aim of our dissertation is to develop an expressive motion detection system to ensure natural interaction between humans and the NAO robot as well as between two NAO robots. For this purpose, in addition to paying attention to human movements while performing a gesture, we also paid attention to the movements and how the gesture is performed by Nao. The idea here is to have a robot capable of recognizing the movement of the person and also his state in an automatic way. The authors in [155, 21], studied the variation of characteristics in the same movement performed with 4 emotions. In our case, we are developing an automatic gesture recognition system that classifies gestures. We consider several gestures made with 4 emotions. Then, we quantify all the components of LMA (body, space, form and effort) and also some spatio-temporal components in order to describe the quantitative and qualitative aspect of the movement. Our movement descriptors are able to differentiate between movements and also the emotions expressed by the same movement. We have built a data set of expressive gestures accessible to the public, easy to use and to be enriched by any person. This data set is made up of 5 expressive gestures. Each gesture is performed with different emotions. We evaluate our motion descriptor on public data set as well as on our constructed data set. Two different feature selection algorithms are developed to study the importance of each movement parameter to discriminate each emotion. A second evaluation of our system is carried out with a human approach based on the opinions of human in the perception of emotions and in the estimation of the proposed descriptor. Finally, to evaluate the robustness and the adequacy of our system, we compare the results of our automatic recognition system with the results obtained from the human approach.

Chapter 3

Recognition of dynamic gestures by Dynamic Time Warping

Contents

In this chapter we present our dynamic gesture recognition system with its different stages, in particular feature extraction and gesture classification.

3.1 Local descriptor inspired by LMA

In order to have a good performance in autonomous action recognition, we seek to build a motion descriptor that is both robust and independent of certain constraints that can influence our gesture recognition system, such as gender or age of the participants. Indeed, a young child, an adult man or an elderly woman will produce the same gesture differently depending on their rhythms, which risks having different representations of the movement. The speed at which a gesture is performed also varies depending on the context and the condition of the person. The same gesture can be performed at a different speed depending on the person's current mood

or the specific intention he/she wants to communicate. So, to be able to have a high accuracy in human gesture recognition system, our descriptor is based on three qualities of LMA: Body, Space and Shape. The Effort component makes it possible to describe the intention, the rhythm and the expressiveness of the movement, whereas this is not the goal of our first application.

3.1.1 Body component

The body component describes the structural and physical characteristics of the human body during movement. In the state of the art, few researchers have used the two components Body and Space, most have focused on Effort and Shape to characterize the quality of movement and describe the expressiveness of the gesture [6, 5, 210, 156, 59, 250]. For our case, this component is essential because we need it to describe the structural characteristics of the body and to distinguish between the different gestures. For the same principle, some researchers have not ignored this factor in their application, [162, 254, 222, 23, 9, 10] who have adopted the LMA approach for analysis of dance gestures, where they quantified the four components of LMA. For the Body, they extracted the following 9 characteristics:

- The distances: between the the knees and the ground, between the hands and the shoulders, between the two hands, between the two feet, between the head and the hands.
- The heights of the pelvis and centroid of the skeleton in relation to the ground.
- The difference between the following two distances: the distance between the hips and the ground and the distance between the feet and the hips.

In the method proposed in [116], which consists in indexing a movement in a very large data set, the authors relied on the components of LMA in order to be able to recover complex movements. They represented the Body factor with the following characteristics:

- A boolean value indicating the presence or absence of motion by comparing the displacement of a segment of the body between two successive frames with a predefined threshold.
- The displacement and orientation of an end effector relative to its root.
- The index of the closest body segment to the end effectors as well as the distance between the two.
- The position of the skeletal center of mass and its displacement in relation to its rest position.
- A Boolean value to indicate the relative position of the center of mass with respect to the body skeleton support polygon.
- An index that identifies the current body part that is used to support the body weight and is the part of the body that is in contact with the ground. Possible values for these indices are: LFoot, RFoot, BothFeet, LHand, RHand, BothHands and can be expanded to include any part of the body if necessary.

Likewise the authors in [250, 251], used the LMA technique to analyze the gestures of a conductor. They quantified this component with three different measures: the distances between the hands and the shoulders

(left and right parts) and the spatial asymmetry of the body. The characteristic of body symmetry has also been used in several works, [86, 269, 81], in order to describe the symmetry of the body during the representation of gestures expressive.

In our work, in order to express the connectivity of the body and find the relationship between the parts of the body, several characters are defined for this component. To do this, we prepared the blue NAO robot to perform several gestures by the Choregraphe software. The purpose of this was to find out what components a humanoid robot uses more to perform various gestures such as dancing (Figure 4.1).

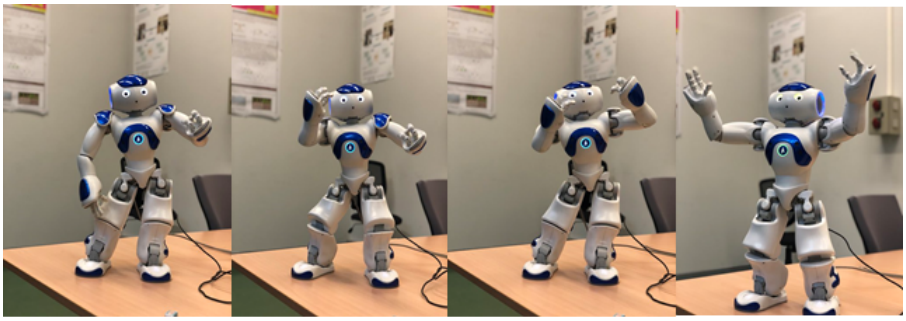


Figure 3.1: The various poses of NAO while dancing.

Since we wanted to create a bi-lateral interaction between two humanoid robots (NAO blue and Nao Orange), we chose the parameters of the NAO's body that have the most movement during the dance pose and are therefore recognizable to the orange robot. As we can see in Figure 4.1, the angles between the joints such as the knees and elbows, the angular rotation of the wrists, the distance between the hands and other parts of the body are the parameters that attract the most attention. Therefore, according to these observations and also considering human movements to create a bi-lateral interaction between humans and robots as well, the following parameters have been selected to construct the descriptor.

The first character defined for the body element is the 3D coordinates of all the joints. Let consider $j_i = (x_i, y_i, z_i)$ is the i_{th} joint of the skeletal representation of the body, so for all joints we will have $J_n = \{j_1, \ldots, j_i, \ldots, j_n\}$, where n is the number of joints and can vary depending on the type of used sensor. The next element intended for the body component is the vector between all relative joints, to do this, we compute all combinations of J_n taken 2 at a time and then the vector between them. We also added angles of body parts to the descriptor. So the angles around elbows (θ_r^1, θ_l^1), neck (θ_r^2, θ_l^2), hips (θ_r^3, θ_l^3) and knees (θ_r^4, θ_l^4) are computed by:

$$\theta^{j_j} = \arccos \frac{\overrightarrow{j_j - j_i} \cdot \overrightarrow{j_k - j_j}}{\left\|\overrightarrow{j_j - j_i}\right\| \left\|\overrightarrow{j_k - j_j}\right\|} \tag{3.1}$$

Where $j_i = (x_i, y_i, z_i)$, $j_j = (x_j, y_j, z_j)$, and $j_k = (x_k, y_k, z_k)$ $3D$ coordinates of three consecutive joints. Figure 3.2 shows an image of the extracted characters in the Body component in a skeletal representation of 15 joints. The *"l"* index represents the left side of the skeleton and the *"r"* index represents the right side of the skeleton. Figure 3.3 shows the θ_l^2 angle changes as well as the position of the left hand for pouring water 3.4. The changes of distances between the left hand and the hip center as well as the left hand and head for the same pose is shown in Figure 3.4.

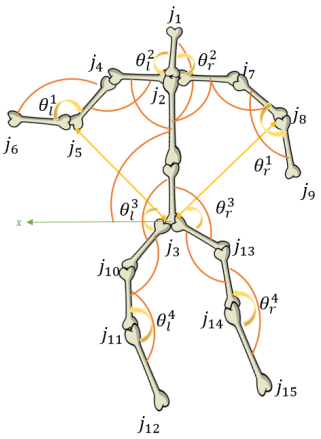


Figure 3.2: Characters extracted from 3D skeleton representation for body component.

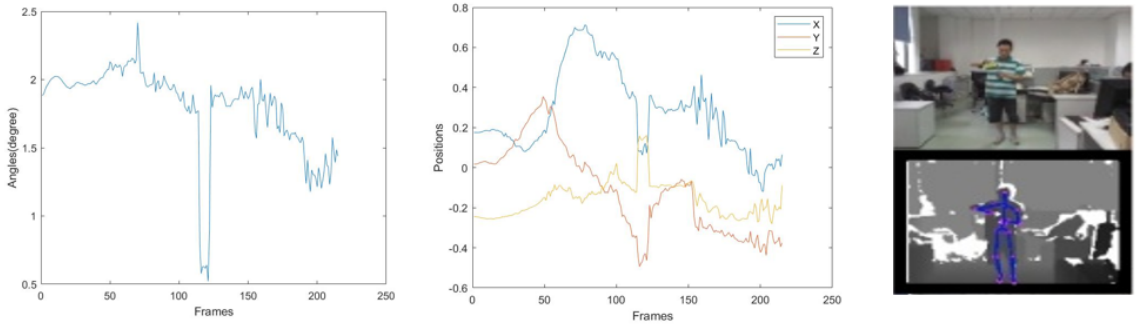


Figure 3.3: Representation of a) Changes in θ_l^2, b) X, Y and Z changes in the left hand and c) Skeleton body during the gesture of pouring water.

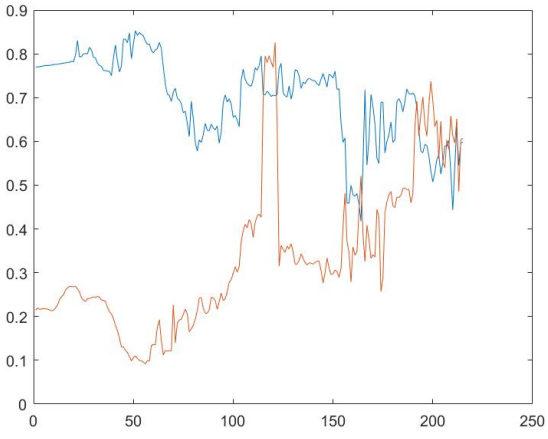


Figure 3.4: The distance between the left hand and the hip center as well as the left hand and head with respect to frames for pouring.

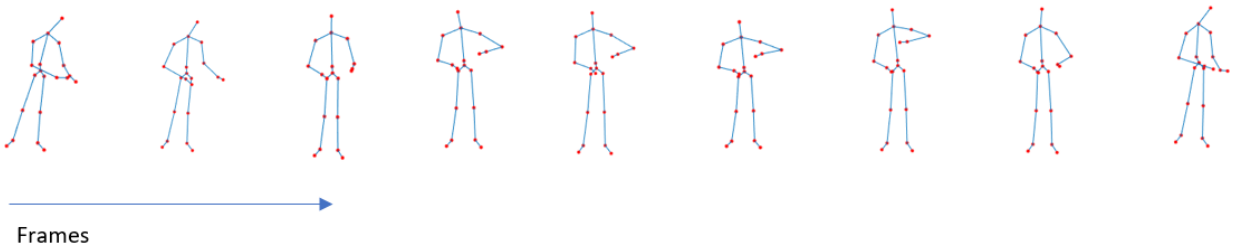


Figure 3.5: Skeletal representation of pouring gesture.

As mentioned, some researchers have categorized movements according to whether they are symmetrical or not. In this case, angles such as θ^4 can indicate the symmetry or asymmetry of a gesture. For example, for the "sitting on chair" gesture 3.7 in the SYSU HOI data set, as shown in Figure 3.6, the variations of these angles are similar.

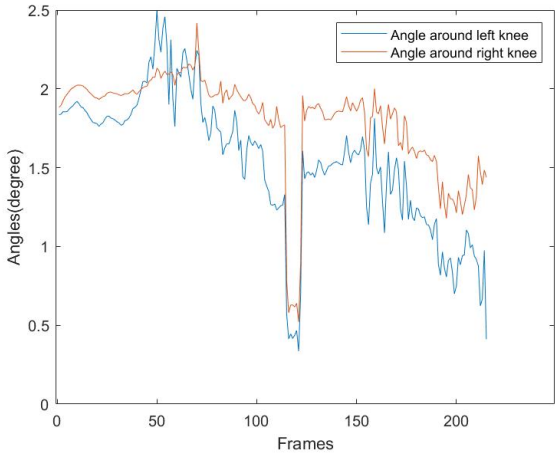


Figure 3.6: The variation of θ^4 for the "sitting on chair" gesture.

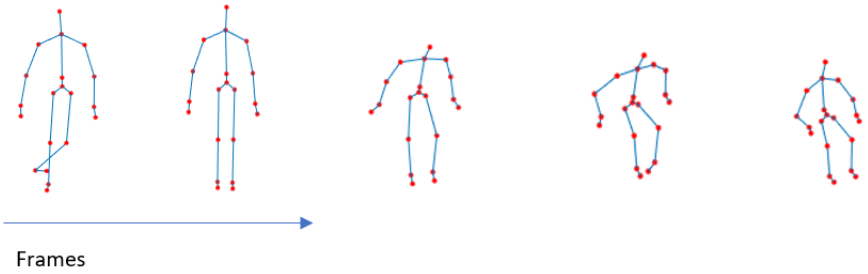


Figure 3.7: Skeletal representation of sitting on the chair gesture.

3.1.2 Space component

The Space component describes the location, directions and paths of a movement. In [250, 251], the authors quantified the quality of Space with two characteristics, the position of the head to characterize the forward backward movement of the head and the forward tilt angle defined as the angle between the vertical axis Y and axis connecting the center of the hip and the head. The authors in [9, 10, 13], also used two different

characteristics to characterize the Space which are the total distance traveled over a period of time where they used it for the evaluation of three different durations of 30, 15 and 5 seconds and the area covered for the same period. In order to quantify this component, we calculated the curve created by hands so that for each frame, f_i, the curve formed by the hand is calculated with the starting point of the f_1 and the end point of f_i. The inner points are a non-linearly row vector of 50 evenly spaced points between f_1 and f_i. So for all $1 \leq i \leq n$ and $1 \leq f \leq F$, where F is the total number of the frames in a motion sequence, we will have:

$$Curve(j_h) = spline(j_h^{f_1}, j_h^{f_i}) \qquad h = left \ \ and \ \ right \ \ hands \qquad 1 \leq f \leq F \tag{3.2}$$

Figure 3.8, shows the curves of this method for the left hand in the "waving" gesture and for the two hands in the "clapping" gesture in the FLORENCE 3D ACTIONS data set.

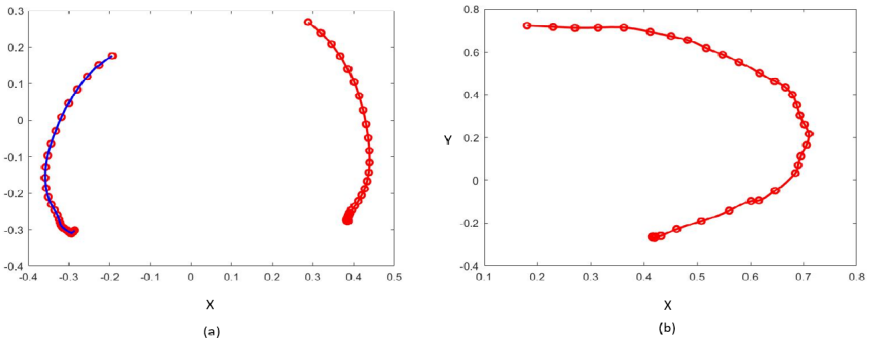


Figure 3.8: Display the curve of a) two hands in ′clapping′ and b) left hand if ′waving′ gesture in FLORENCE 3D ACTIONS data set.

Another sub-component of the component of space that can help make this descriptor more robust is geometrical observations, whose task is to describe a movement in terms of its direction and location in its environment. For this sub-component, we calculated the quaternion of all the angles shown in red in Figure 3.2 in the local coordinates.

3.1.3 Shape component

The Shape component involves three distinct qualities of change in the form of movement: shape flow, directional movement, and shaping.

3.1.3.1 The shape flow

This factor reflects the relationship of the body with itself. The changes can be seen as the increasing or decreasing volume of the shape of the body. Some authors have represented the Shape component with only this factor and have ignored the others, such as [9, 10, 13, 116, 156, 210] who used the limiting volume of the skeleton as a measure for this factor. They divided the skeleton into 4 parts (upper, lower, right and left) and calculated their corresponding volumes and the limiting volume of all the joints. In addition, they added two measures, the height of the torso (the distance between the head and the center of the hip) and the level of

the hands (the level higher above the head, the intermediate level between the head and the midpoint between the head and the center of the hip, and the low level below the midpoint). The authors in [221], calculated the The volume of the 5 joints(head, hands, and feet), volume of upper body, volume of lower body, volume of left and right sides, torso height and hands level as the shape factor. The authors in [86], calculated the area of a triangle connecting the following three joints: the head, the right hand and the left hand. The authors in [251], quantified the flow of form by an index relating to the contraction of the body in order to characterize the extension of the limbs in relation to the center of the body. In this work, for the five joints that have the highest degree of freedom of movement, respectively: head, left hand, left foot, right foot, and right hand, we calculated the volume of polyhedron created by these joints. This step is done by calculating the volume of the convex hull of the 3D skeleton based on Quickhull algorithm [22]. This component helps us to interpret how a movement's shape changes, so we can determine whether the occupied space by the body, increases or becomes narrower. Figure 3.9 shows the polygonal changes of these joints in the "Clapping" gesture over time.

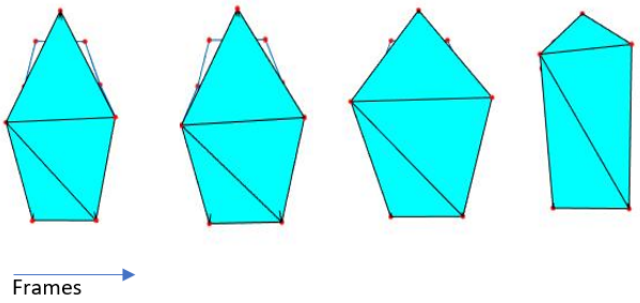


Figure 3.9: The variation of Polygonal consisting of head, left hand, right hand, left foot and right foot.

Finally, for a data set consisting of N_v videos, in which each skeleton, according to the type of used sensor, has n joints, the descriptor, F, is as follows:

$$F = \left[M_{m\times D}^{1,1,1} \quad M_{m\times D}^{1,1,2} \quad \dots \quad M_{m\times D}^{2,1,1} \quad \dots \quad M_{m\times D}^{a,p,r} \quad \dots \quad M_{m\times D}^{N_a,N_p,N_r} \right]_{N_v\times m\times D}^{T} \tag{3.3}$$

Where T represents transpose, $N_v = N_a \times N_p \times N_r$ and for each $1 \le a \le N_a$ (Number of actions), $1 \le p \le N_p$ (Number of participants) and $1 \le r \le N_r$ (Number of repetition), we have:

$$M_{m\times D}^{a,p,r} = \left[v^1 \dots v^D \right]_{m\times D} \tag{3.4}$$

where for all $1 \le d \le D$(number of desired frames)

$$v^d = \left[v_{Position}^{1\times(3\times n)} \quad v_{ArcLenght}^{1\times(n-1)} \quad v_{VolumOfPolyhedron}^{1\times 1} \quad v_{RelativeJoints}^{1\times(3\times C_2^{J_n})} \quad v_{ThetAngles}^{1\times 8} \right]_{1\times m}^{T} \tag{3.5}$$

where $[]^T$, indicates the transpose of a matrix and $C_2^{J_n}$ is equal to the number all combinations of J_n taken 2 at a time and $m = (3\times n) + (n-1) + 1 + (3\times C_2^{J_n}) + 8$.

As mentioned above, the LMA algorithm consists of four components. The fourth component, or Effort, relates to how to perform a movement that deals with the speed, acceleration, or force used to perform a gesture. Since

we only deal with gestures in this section, and the way they are handled by the participants and their moods or emotions is not discussed here, we skip this component. Because if in the case of including elements such as speed, there will be a difference between a slow-moving and a fast-moving one, and the identification accuracy will decrease.

3.2 Dynamic Time Warping

DTW is one of the common methods for measuring the similarity between two different curves. This optimization algorithm can compress or stretch the signals adaptively to create an optimal map between two time series. Using the normalized path to calculate the similarity between sequential data can overcome the problem that temporal data cannot match each other because of different signal lengths[292]. In the following, we will first review the applications of this algorithm and then explain how it works.

3.2.1 The applications of Dynamic Time Warping

Dynamic Time Warping (DTW) algorithm is a well-known and popular algorithm in many fields. This algorithm was first introduced in the 1960s [28] and has been widely studied and tested. It was used in the 1970s in speech recognition filed in[172] and [209] and is now used in many fields: handwriting and online signature matching [183, 236, 238], sign language recognition [207, 195, 241] and gestures recognition [216, 54, 27], data mining and time series clustering [141, 122, 198, 55, 149], computer vision, and computer animation [169], video surveillance system [121, 180], Protein Sequence similarity and Chemical processing [100, 60], Music and Signal Processing [190, 178, 194, 110, 18, 71].The author in [192], developed an interface in order to track the hands of participants by the use of Kinect sensor in real time and identify the gesture performed by them by the use of DTW. Their proposed framework is shown in figure 3.10.

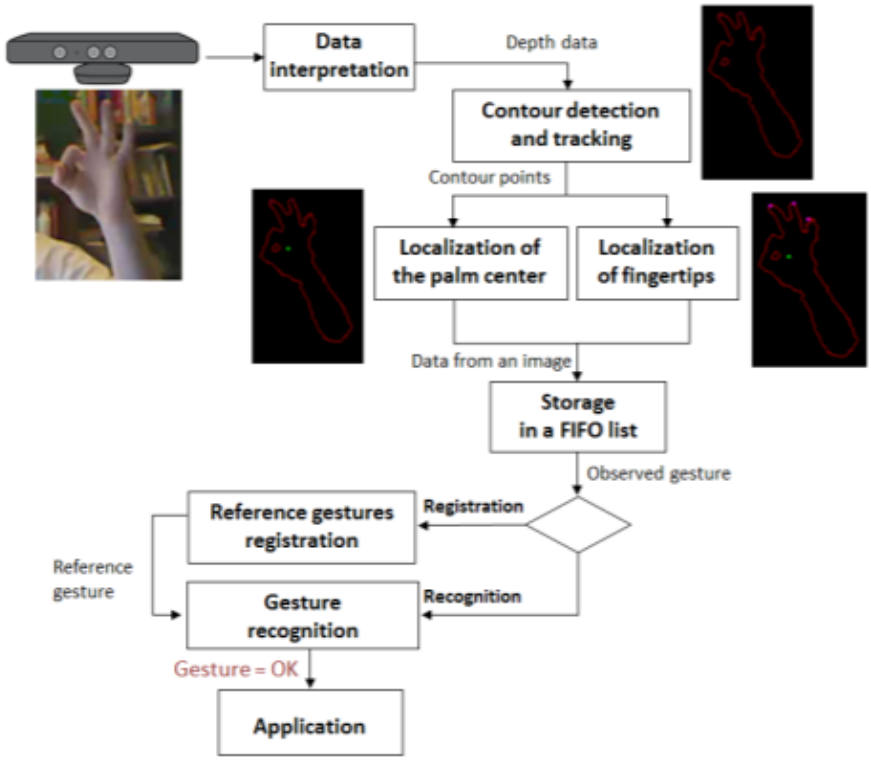


Figure 3.10: Proposed framework for hand detection and gesture recognition.

In [167], the authors present an innovative average of a set of time series based on DTW. In this method, so called CDBA (Constrained DTW Barycenter Averagingin), in addition to the mean value, another variable called tolerance in each time step is proposed to model the time series changes in the mean range. In order to evaluate the proposed method, it was implemented on e UCR time-series Classification Archive data set and the results were in many cases better than the results in the literature. In [271], the authors use Time Weighted DTW (TWDTW), to analyze time series satellite images. To do this they designed a package so called R dtwSat. Their proposed package consist of several functions to:

- Build temporal patterns for land cover types by the use of TWDTW analysis through different weighting functions.
- Simulate the results in a graphical interface.
- construct land cover maps.
- build spatio-temporal plots for land changes

Figure 3.11 shows an example of the simulated results and constructed land map cover generated by this package. In the following, we will review how Dynamic Time Warping works and its application in this study.

3.2.2 The algorithm of Dynamic Time Warping

The DTW method has gained its popularity by using very high performance as a method of measuring time series similarity. This algorithm minimizes the effects of the changes and distortions over time series data by allowing "elastic conversion" of the time series in order to identify similar warped curves with different steps.

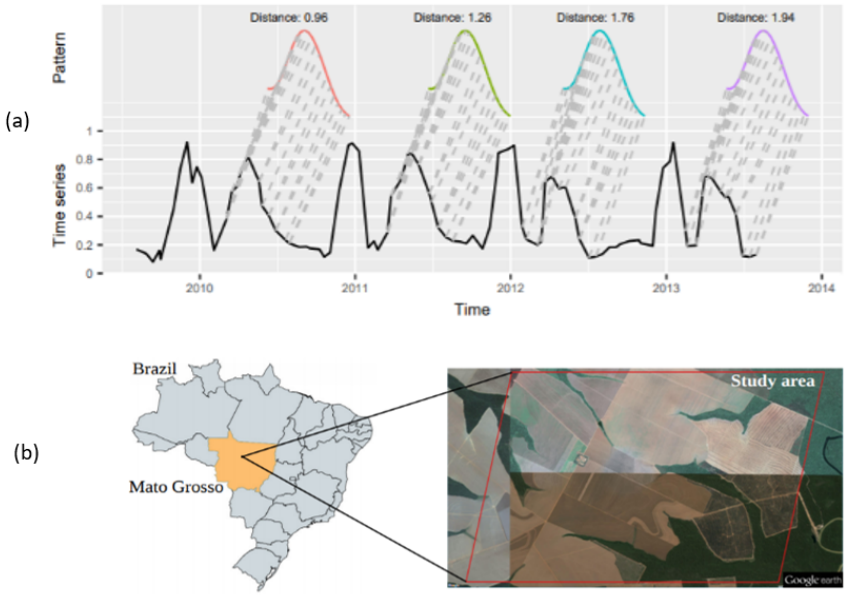


Figure 3.11: Representation of a: Four best matches of the "soybean" pattern in the time series using a logistic time-weight. The solid black line is the long-term time series; the colored lines are the temporal patterns; and the gray dashed lines are the respective matching points and b: Constructed land cover maps.

This optimization algorithm can compress or stretch the signals adaptively to create an optimal map between two time series. Using the normalized path to calculate the similarity between sequential data can overcome the problem that temporal data cannot match each other because of different signal lengths[292]. For two time series, $X = (X_1 \dots X_N)$, $N \in \mathbb{N}$ and $Y = (Y_1 \dots Y_M)$, $M \in \mathbb{N}$ their warp path can be expressed as $w_1, \dots, w_k, \dots, w_K$, $(max(N,M) \leq K \leq N+M)$. Where $w_k(a,b)$ is a link between X_a, $1 \leq a \leq N$, and Y_b, $1 \leq b \leq M$. DTW optimizes the solution in $O(MN)$ time, which can be ameliorated by the use of different methods such as multiple scales [170, 170, 264]. There is a limit to using this method, which is a problem with the data sequence, which means that they must be sampled at convergence points at fixed times (this problem can be solved by re-sampling). If the sequences receive their value from some feature space θ, then in order to compare these two sequences $X, Y \in \Theta$, it is needed to use a local distance measurement, which is a defined as a function:

$$f : \Theta \times \Theta \rightarrow \mathbb{R} \geq 0 \tag{3.6}$$

Obviously, if the sequences are very similar, f has a small value, and if they are completely different, it has a large value. This function is known as the cost function and its task is to rearrange the sequence points by minimizing the cost or distance function. The algorithm starts by constructing an euclidean distance matrix $D \in R^{N \times M}$ whose elements represent the paired distances between X and Y. This matrix, which is used to arrange the two sequences X and Y to find the minimum cost, is called the local cost matrix.

$$C_l \in R^{N \times M} : c_{i,j} = \left\| x_i - y_j \right\|, i \in [1:N], j \in [1:M] \tag{3.7}$$

Once the local cost matrix is created, the minimum path algorithm begins to find the path with the minimum cost. This warping path (or warping function) defines the correspondence of an element $x_i \in X$ to $y_j \in Y$ following three boundary condition which allocate first and last elements of X and Y to each other, Figure 3.12.

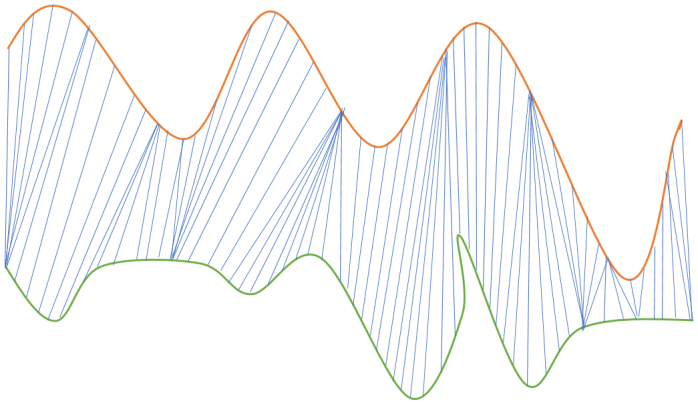


Figure 3.12: The optimal warping path finding the minimum distance between two time series X and Y.

In general, the warped path constructed by DTW is a sequence of points $w = (w_1, w_2, \ldots, w_K)$ with $w_l = (w_i, w_j) \in [1:N] \times [1:M]$for $l \in [1:K]$ which must meet the expectations of following criteria:

- Boundary constraint: $w_1 = (1,1)$ and $w_K = (N,M)$. That means the first and last points of the warped curve must be the first and the last points of arranged sequences.
- Monotonicity constraint: Given $w_k(a,b)$ and $w_{k+1}(a',b')$ then $a \leq a'$ and $b \leq b'$. This status holds the time-ordering of points.
- Continuity constraint: Given $w_k(a,b)$ and $w_{k+1}(a',b')$ then $a' \leq a+1$ and $b' \leq b+1$. This condition restricts the warping curve from long jumps while arranging sequences.

Finally the desired cost function is defined as:

$$c_W(X,Y) = \Sigma_{l=1}^{K} c(x_l, y_l) \tag{3.8}$$

The warping path which has the lowest cost associated with sorting is known as the optimal-warping path. Given the definition of the optimal-warping path in order to find the best one, we must test each possible warping path between X and Y. This goal can be computationally challenging due to the exponential growth of the number of optimal paths resulting from the exponential growth of the length of X and Y. The algorithm used in DTW, to solve the problem of complexity, is DTW distance function which has a linear complexity of $O(MN)$.

$$DTW(X,Y) = c_{w^*}(X,Y) = min\{c_W(X,Y), W \text{ is an } (N,M) - \text{warping path}\} \tag{3.9}$$

So that:

- First row: $D(1,j) = \Sigma_{k=1}^{j} c(x_1, y_k), j \in [1:M]$.
- First column: $D(i,1) = \Sigma_{k=1}^{i} c(x_k, y_1), i \in [1:N]$.
- Other elements: $D(i,j) = min\{D(i-1,j-1), D(i-1,j), D(i,j-1)\} + c(x_i, y_j), i \in [1:N], j \in [1:M])$.

Thereupon, using the DTW [161], we get the total cost of the optimal warping path, W^*. The elements of W^*, are the vector indicators, v_d, belonging to Matrix $M_{m \times D}^{a,p,r}$, that we have to go through, respectively, to

achieve the minimum cost using DTW. In the next step, the matrix $M_{m\times D}^{a,p,r}$ will be updated by replacing its vector with concatenated vectors associated with the indicators in I_Y (Algorithm 1). Afterward, in order to reduce complex time series data, as it is proposed in [166], a Fast Fourier Transform (FFT) (appendix B) is applied to all curves obtained by the DTW. According to the [108], SVM (appendix A) is a powerful classifier for classifying FFT-based data. In general, SVM is a binary classifier, but it can also be used as a multi-class classifier. LIBSVM [52], which is a most widely used tool for SVM classification is implemented to train and label data. The performance of a multi-class SVM method is dependent on the choice of its kernel function. In this paper we employed linear kernel function. linear kernel gives best results by tuning its parameters. For kernel function is defined as: $K(x_i, x_j) = (\gamma x_i^T x_j + a_0)$, where x_i and x_j are vectors of features computed from training instances. γ and a_0 indicate weights and bias, There is also another parameter, penalty parameter C, which is using to reduce the number of training errors. The selected parameters for these values are: $\gamma = 1$, $a_0 = 1$ and $C = 1$. Figure 3.13 is an overview of the steps performed in the proposed pipeline, which we will explain in detail below. The process of converting raw data into a label for each gesture can be divided into the following steps:

1. Extract the LMA features from normalized data, F.
2. Divide F into two categories: $F_{Train} = \begin{pmatrix} M_{m\times D}^{a_{tr}^1,p_{tr}^1,r_{tr}^1} \\ \vdots \\ M_{m\times D}^{a_{tr}^h,p_{tr}^h,r_{tr}^h} \\ \vdots \\ M_{m\times D}^{a_{tr}^{S_{train}},p_{tr}^{S_{train}},r_{tr}^{S_{train}}} \end{pmatrix}$ and $F_{Test} = \begin{pmatrix} M_{m\times D}^{a_{te}^1,p_{te}^1,r_{te}^1} \\ \vdots \\ M_{m\times D}^{a_{te}^b,p_{te}^b,r_{te}^b} \\ \vdots \\ M_{m\times D}^{a_{te}^{S_{test}},p_{te}^{S_{test}},r_{te}^{S_{test}}} \end{pmatrix}$
   In which S_{train} and S_{test} indicate the total number in training and validation data.
3. Select $M_{m\times D}^{a_{tr}^1,p_{tr}^1,r_{tr}^1}$ as a reference matrix and calculate the cost function and warped path between $M_{m\times D}^{a_{tr}^2,p_{tr}^2,r_{tr}^2}$ and reference curve.
4. Update the columns of $M_{m\times D}^{a_{tr}^2,p_{tr}^2,r_{tr}^2}$ according to the warped path and horizontally concatenate the updated matrix with reference matrix.
5. Repeat Step 3 and 4 for all $M_{m\times D}^{a_{tr}^h,p_{tr}^h,r_{tr}^h}$ where $3 \le h \le S_{train}$ to construct:
   $F_{Train}^{Update} = \left[M_{m\times D}^{a_{tr}^1,p_{tr}^1,r_{tr}^1} \dots M_{m\times D}^{a_{tr}^h,p_{tr}^h,r_{tr}^h} \dots M_{m\times D}^{a_{tr}^{S_{train}},p_{tr}^{S_{train}},r_{tr}^{S_{train}}} \right]$.
6. Get the standard deviation of each cell going through the 3rd dimension, h and consider it as a new reference curve.
7. Repeat step 3 to 6 till there no difference between the new reference curve with the previous one.
8. Implement FFT to the warped path.
9. Train the new F_{Train} using linear kernel SVM.
10. Repeat step 3 to 8 for F_{Test}.
11. Label each instance of processed F_{Test} using One Against All SVM.

Algorithm 1 Algorithm for calculating the curve of LMA features using DTW.

1. Input: F
2. Output: Action labels
3. # Divide F into two parts, training, F_{train} and validation, F_{Test}.
4. For $1 \leq a \leq N_a$
5. For $1 \leq p \leq N_p$
6. If $M_{m\times D}^{a,p,r} \in F_{Train}$
7. $F_{Train} \leftarrow M_{m\times D}^{a,p,r}$
8. Else If
9. $F_{Test} \leftarrow M_{m\times D}^{a,p,r}$
10. End If
11. End
12. $ReferenceCurve \leftarrow$ The first $M_{m\times D}^{a,p,r}$ in F_{Train}
13. $Repetition \leftarrow 0$
14. While ($Repetition \leq$Threshold) ► The threshold varies according to the data sets.
15. For $1 \leq h \leq s_{train}$ ► s_{train} = the total number of $M_{m\times D}^{a,p,r}$ in F_{train}.
16. $Cost_{min}, I_X, I_Y \leftarrow$DTW$_{element-wise}(ReferenceCurve^T, F_{Train}(h)^T)$
17. For $1 \leq d \leq D$
18. $Curve(h) \leftarrow$ Replace the d_{th} column of $F_{Train}(h)$ with $v^{I_Y(d)}$
19. End
20. End
21. $RefrerenceCurve \leftarrow$ standard deviation of the elements of $Curve$ along S_{train}.
22. $Repetition \leftarrow Repetition + 1$
23. End
24. $Curve \leftarrow$ FastFourierTransforms($Curve$)
25. $F_{Train} \leftarrow Curve$
26. End
27. Train the F_{Train} using linear kernel Support Vector Machine.
28. Repeat steps 12 to 26 for F_{Test} to obtain the validation matrix, M_{test}.
29. Label samples in F_{Test} using All Against One Support Vector Machine.

These steps are briefly organized in Algorithm 1:

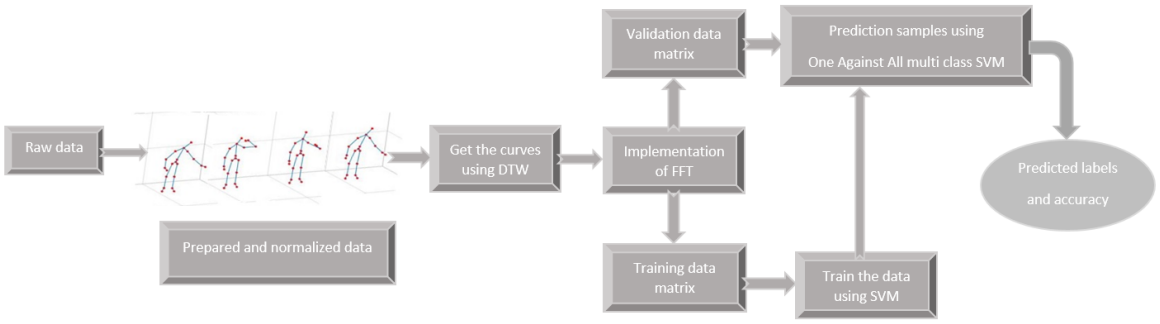


Figure 3.13: An overview of all the steps involved in Algorithm 1 .

3.3 Experiment and Results

In this section, we evaluate the proposed descriptor that we defined using the LMA method on four public data-sets using Dynamic Time Warping algorithm.

As mentioned above, the proposed method in this chapter has been evaluated on four general data-sets namely MSR Action 3D, UTKinect-Action3D, Florence 3D actions and SYSU 3D HUMAN-OBJECT INTERACTION data-sets. The software used to produce the experimental results is MATLAB 2018b.

Before showing the results, a brief description of each data set is given below.

3.3.1 Development and Result

Normalization The first step is to normalize raw data. Normalization of data is important because an action in a data set is performed by different people of different sizes and also in different primary locations. So in the case of non-normalization, the proposed descriptor which is based on the three-dimensional coordinates of the joints and also vectors between successive joints and their size, cannot be invariant, and this can have an effect on reducing their robustness, which reduces gesture recognition accuracy. To make the skeletal data invariant to the body size of the subjects, each component of all vectors between two consecutive joints is divided by its magnitude. Also, in order to make the skeleton data invariant to the angle of each person relative to the used sensor (which is Microsoft Kinect Sensor in all four data-sets used in this work), we matched the X axis of hip center to the X axis of Kinect axis using the rotation matrix.

Data Sampling The next step in preparing the data for use in the algorithm described in the previous section is to fix the number of frames per gesture. For this purpose, the data should be sampled according to the number of desired frames. To do this, the spline function is used. As shown in the figure 3.14, when the number of frames is fixed using the spline function, there is no change in the overall motion pattern.

Angle Presentation In the last step, in order to evaluate the efficiency of the proposed quaternion angles, we calculated the triple-angle of Roll, Pitch and Yaw. As shown in figure 3.15, an action preformed by two different participants has the same amplitude variations.

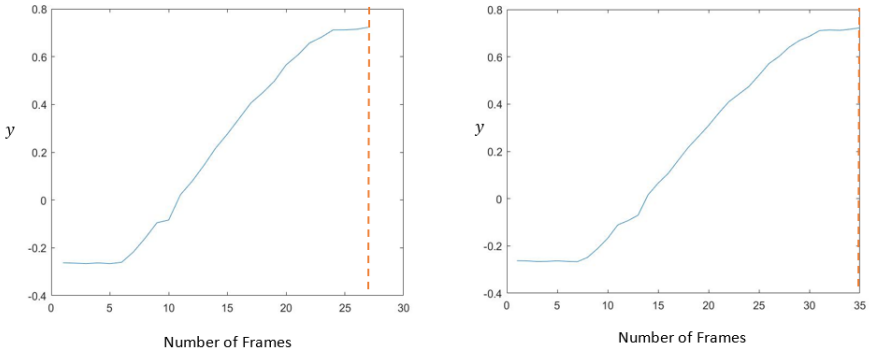


Figure 3.14: Demonstration of hand movement on the Y axis in the Waving gesture with a) actual number of frames, b) desired number of requested frames.

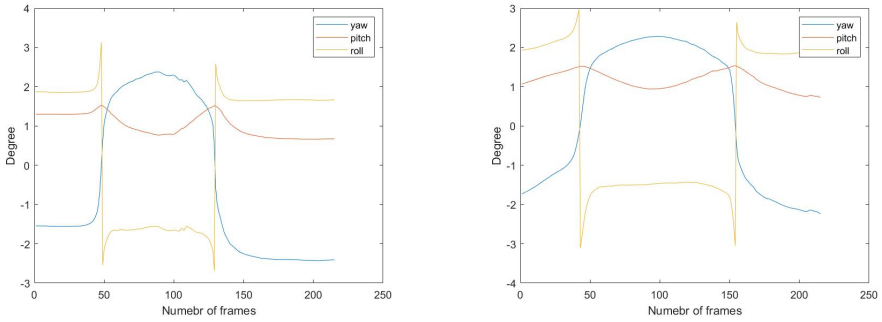


Figure 3.15: (a) Yaw, Pitch and Roll Changes for Person30, Action7; (b) Yaw, Pitch and Roll Changes for Person40, Action7.

3.3.2 Comparison With The Sate of Art and Discussion

According to the state of art, there are two methods to use the MSR-Action 3D data-set. In the first one, the data-set is used as a whole, we used 60% of the data for training and 40% for testing whereas the data are randomly divided. Using this method, the calculated average accuracy is 90.55%, as shown in the table 3.1, this value is superior to similar work in the field of action recognition based on skeletal data. The confusion matrix of this setting is shown in figure 3.16.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	0.97	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
3	0.00	0.00	0.54	0.00	0.12	0.03	0.03	0.05	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.16	0.00	0.00	0.10	0.32	0.12	0.07	0.00	0.00	0.00	0.00	0.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
5	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
6	0.05	0.00	0.00	0.07	0.00	0.88	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
7	0.06	0.11	0.00	0.00	0.00	0.00	0.70	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
9	0.05	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.83	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.99	0.00	0.00	0.00	0.00	0.00	0.00	0.01
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.99

Figure 3.16: Confusion matrix of MSR-Action3D data-set.

Due to the existence of large amount of computation in training and validation phases that may lead to decrease the accuracy, the author in [145], have proposed to divide the whole data-set into three different action

Method	Year	Accuracy(%)
Active Joints [246]	2107	84.72
EigenJoints[287]	2012	82.30
AHON4D [179]	2013	88.36
Cooperative Warp [239]	2019	90.90
Actionlet Ensemble [262]	2012	88.20
Coding Kendall's Shape Trajectories [29]	2018	86.18
Learning Composite Latent Structures[272]	2019	87.2
HAR using CNN[4]	2019	87.1
Our method	2020	90.55

Table 3.1: Comparison with the state-of-the-art results MSR-Action3D data-set.

categories (AS1,AS2,AS3) (Table 3.2), each consisting of 8 actions.

Action Set 1(AS1)	Action Set 2(AS2)	Action Set 2(AS2)
Horizontal arm wave	High arm wave	High throw
Hammer	Hand catch	Forward kick
Forward punch	Draw x	Side kick
High throw	Draw tick	Jogging
Hand clap	Draw circle	Tennis swing
Bend two	Hand wave	Tennis serve
Tennis serve	Forward kick	Golf swing
Pickup and throw	Side boxing	Pickup and throw

Table 3.2: Subsets of actions, AS1, AS2, and AS3 in the MSR Action 3D dataset.

By implementing the second method, the recognition accuracy has reached 99.24 (Average accuracy of AS1, AS2 and AS3). Given that the size of each category is small, we used only 30% of the data for validation, and 70% of it was allocated to the training part. In AS1 and AS2, most movements are related to the upper and middle trunk, and in some cases the movements are very similar, for example drawing x and ticking. Therefore, it is reasonable to have a low accuracy compared to AS3, which consists of gestures in which all body organs are involved and are also very different. The results of this method and comparison with previous work can be seen in the table 3.3. And the average confusion matrix of AS1, AS2 and AS3 is shown in figure 3.17. The

Method	Year	AS1 Accuracy(%)	AS2 Accuracy(%)	AS3 Accuracy(%)
Mining Key Skeleton Poses with Latent SVM[146]	2017	89.1	88.7	94.9
Lie Group using deep network [202]	2018	96.64	87.52	98.71
LMA Qualities [6]	2019	90.3	88.7	93.1
Improving bag-of-poses[1]	2019	94.3	94.6	97.7
Coding Kendall's Shape Trajectories [29]	2018	95.87	86.72	100
DMM-UDTCWT [239]	2019	95.6	93.82	96.6
Our method	2020	99.13	98.60	100

Table 3.3: Comparison with the state-of-the-art results AS1/AS2/AS3.

results obtained from the UTKinect Action, 60% training and 40% testing, Florence 3D and SYSU 3D HOI data-sets, 60% training and 40% testing, respectively, are as follows (Table 3.7). As we can see, our proposed method outperforms most of the work in the state of art.

Figure 3.18, represents the confusion matrix of UTKinect Action data-set.

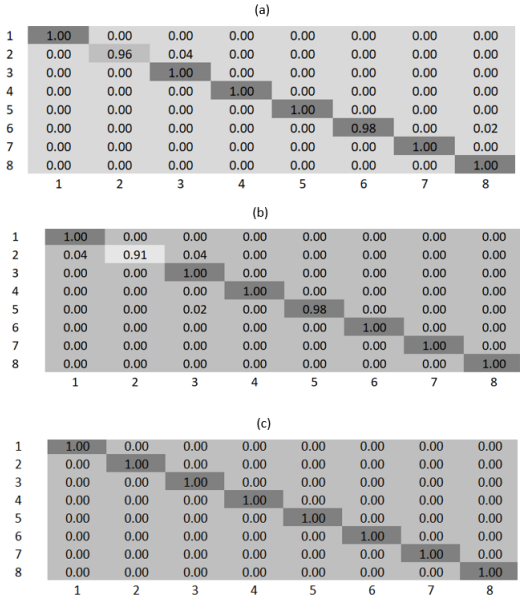


Figure 3.17: Confusion matrix of sub categories of MSR-Action3D data-set: a) AS1, b) AS2 and c) AS3.

Method	Year	UTKinect Action Acc(%)
Active Joint [246]	2017	95.9
Mining Key Skeleton Poses with Latent SVM[146]	2017	91.5
Motion Trajectories [64]	2012	91.50
Cooperative Warp [239]	2019	95.38
Grassmann manifold [231]	2015	88.50
Group Sparse Regression[143]	2018	95.1
Traj. on $S^+(3,n)$- BP Fusion[115]	2018	96.48
SVRNN[41]	2019	89.0
Our method	2020	97.36

Table 3.4: Comparison with the state-of-the-art results UTKinect Action.

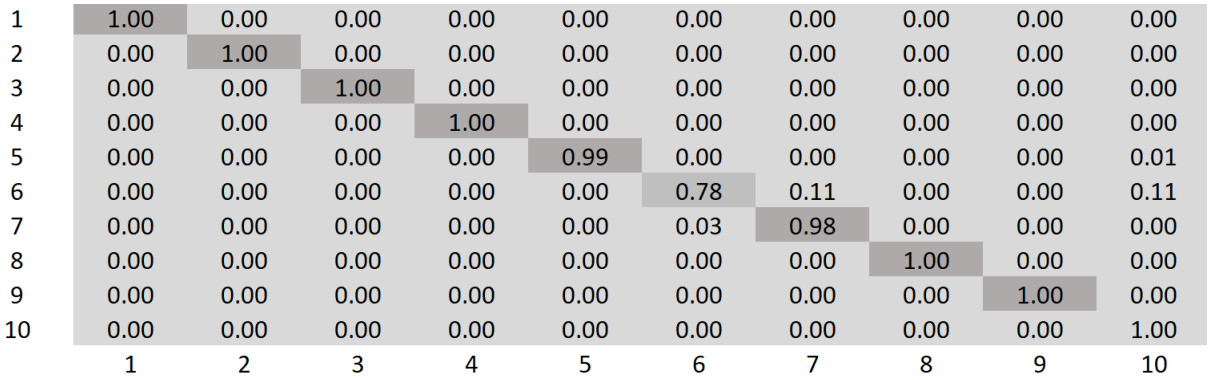


Figure 3.18: Confusion matrix of UTKinect Action data-set.

For the SYSU data set, in addition to the general mode of its use (Setting 1), considering the entire data set, since its size ois the same as MSR-Action3D, we decided to divide this data set into different sections (Setting 2), just like MSR-Action3D. In this data set, two movements are performed with each object. In order to divide, we decided to divide it into two parts so that in each part, only one movement is done with each object (Table 3.6). Drinking, Calling phone, Wearing backpacks, Sitting chair, Taking out wallet and Mopping are placed in the first category (AC1), and Pouring, Playing phone, Packing backpacks, Moving chair, Taking from wallet and Sweeping are in the second category (AC2). After implementing this division, allocating 40% of the data

for validation and 60% for training, the accuracy increased to **92.32**% (accuracy of AC1 (**92.31**%) and AC2 (**92.33**%)). The result is 5% better than the setting 1, without splitting the data.

Method	Year	SYSU 3D HOI Acc(%)
Group Sparse Regression[143]	2018	80.7
Reinforcement Learning[244]	2018	76.9
Self-Attention Guided Deep Features[281]	2019	80.36
HRS networks[282]	2019	84.23
Traj. on $S^{+}(3,n)$- BP Fusion[115]	2018	80.22
Physiological function assessment [50]	2019	83.75
Geometric Algebra Representation [49]	2019	84.62
Progressive Teacher-student Learning [270]	2019	87.92
Deep Bilinear Learning [103]	2018	88.9
Our method	2020	Setting 1:86.63/2: **92.32**

Table 3.5: Comparison with the state-of-the-art results SYSU 3D HOI.

Action Category 1(AC1)	Action Category 2(AC2)
Drinking	Pouring
Calling phone	Playingphone
Wearingbackpacks	Packing backpacks
Sitting chair	Moving chair
Taking out wallet	Taking from wallet
Mopping	Sweeping

Table 3.6: Subsets of actions, AC1 and AC2 in the SYSU 3D HOI data set.

Finally, Table 3 compares the results obtained from the implementation of the proposed method on the Florence data set and Figure 5 shows its matrix.

Method	Year	Florence 3D Acc(%)
Mining Key Skeleton Poses with Latent SVM[146]	2017	87
Motion Trajectories [64]	2012	87.0
Cooperative Warp [239]	2019	88.38
Graph-Based[265]	2016	91.63
Lie Group [258]	2014	90.88
Intrinsic SCDL (Bi-LSTM) [243]	2019	93.04
Transition-Forest [82]	2017	94.16
SCK+DCK [133]	2016	95.23
Our method	2020	94.22

Table 3.7: Comparison with the state-of-the-art results Florence 3D.

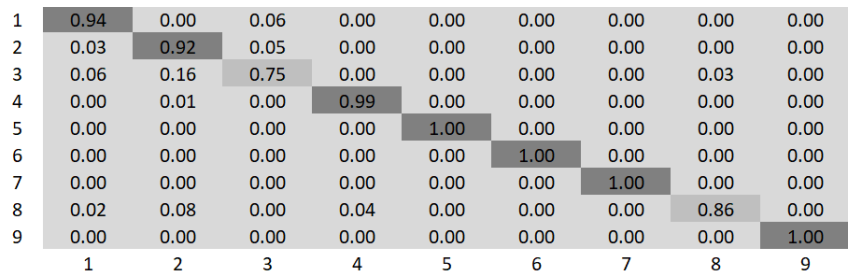


Figure 3.19: Confusion matrix of Florence 3D-Action data-set.

3.4 Conclusion and Future Work

In this section, we have introduced a robust and invariable descriptor using the Laban Movement Analysis method, since our focus was only on gesture recognition, so only three of the four components of this method were used. In this approach, elements must be selected that reflect the changes in the body as it moves. For example, one of the elements added to this work is the angle around the hip. Since the data-sets tested here often include gestures that require all parts of the body to perform, this angle, is a good linkage for connecting the upper and the lower trunk. Also, in this chapter, it was shown that by dividing the SYSU 3D HOI data-set into two parts (AC1 and AC2), accuracy can be increased for 5.69%. In the second part, the performance of the proposed descriptor by the use of dynamic time warping method was tested. Four public data-sets were investigated. The obtained results in our method were in many cases superior to other studies in gestures recognition based on 3D skeletal coordinates. The lowest accuracy rate was related to the MSR Action 3D data-set, which, after dividing it into three groups, revealed that there were many similar movements in one group ($AS2$), and that the same movements in the overall data set caused confusion. In the next section, we use the deep learning method to identify gestures in order to evaluate large data-sets.

Chapter 4

Autonomous Gesture Recognition using Multi-layer LSTM Networks and Laban Movement Analysis

Contents

In recent years, due to the reasonable price of RGB-D devices, the use of skeletal-based data in the field of human-computer interaction has attracted a lot of attention. Being free from problems such as complex backgrounds as well as changes in light is another reason for the popularity of this type of data. In the existing methods, the use of joints and bone information has had significant results in improving the recognition of human movements and even emotions. However, how to combine these two types of information in the best possible way to define the relationship between joints and bones is a problem that has not yet been solved. In this chapter, we use a new method based on the geometric coordinates of different parts of the body to show a movement. To do this, in addition to the distances between hip center and other joints of the body and the changes of the quaternion angles in time, we define the triangles formed by the different parts of the body and calculated their areas. We also calculated the area of the single conforming 3-D boundary around all the joints of the body. At the end we added the velocity of different joint in the proposed descriptor. We use a long short-term memory (LSTM) network to evaluate this descriptor. The proposed algorithm is implemented on

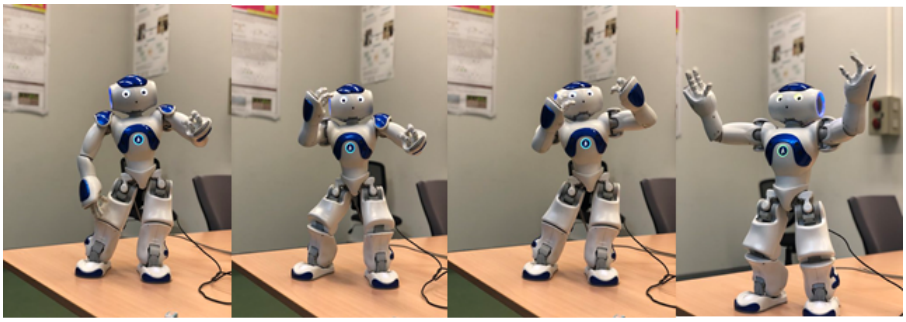


Figure 4.1: The various poses of NAO while dancing.

two general data sets, NTU RGB+D 120, SYSU 3D HOI, FLORENCE 3D 95.24, ACTIONS DATASET and the results are compared with those available in the literature.

4.1 Proposed Feature Descriptor

Construction of a precise and robust descriptor is one of the most important steps in the process of pattern recognition. When dealing with the three-dimensional data of the joints and the information obtained from the skeleton of the working body, we must pay attention to the fact that this data can change according to different conditions. For example, the three-dimensional coordinates of the joints varies with the changes of the initial position and rotation of the person performing the gestures, according to the global frame. Therefore, without normalization of data, the 3D coordinates of the joints can not be reliable data for recognizing a gesture. To this end, we have selected factors that are immutable to these factors. Factors that are variable in relation to these factors were also normalized. In the following of this section, we will first define LMA and then we will detail the construction of a robust descriptor by the use of this algorithm.

Laban Movement Analysis (LMA) [19] is an algorithm to analyze, visualize and describe the human gestures and emotions using its four components, namely body, shape, space and effort which are composed of spatio-temporal features.

Since in this thesis we are dealing with gestures without emotions, then in this study, we investigate the first three components that means Body, Shape and Space. The body component describes the structural and physical characteristics of the human body during movement. In this paper, in order to express the connectivity of the body and find the relationship between the parts of the body,several characters are defined for this component. To do this, using Choregraphe we prepared the blue NAO robot to perform several gestures. The purpose is to find out what components a humanoid robot uses more to perform various gestures such as dancing (Fig 4.1).

The first character defined for the body element is the 3D coordinates of all the joints. Let consider $j_i = (x_i, y_i, z_i)$ is the i_th joint of the skeletal representation of the body, so for all joints we will have $J_n = \{j_1, ..., j_i, ..., j_n\}$, where n is the number of joints and can vary depending on the type of used sensor. The second sub components of body are the angles of body parts to the descriptor. So the angles around elbows (θ_r^1, θ_l^1), neck (θ_r^2, θ_l^2), hips (θ_r^3, θ_l^3) and knees (θ_r^4, θ_l^4) are computed by:

$$\theta^{j_j} = \arccos \frac{\overrightarrow{j_j - j_i} \cdot \overrightarrow{j_k - j_j}}{\left\|\overrightarrow{j_j - j_i}\right\| \left\|\overrightarrow{j_k - j_j}\right\|} \tag{4.1}$$

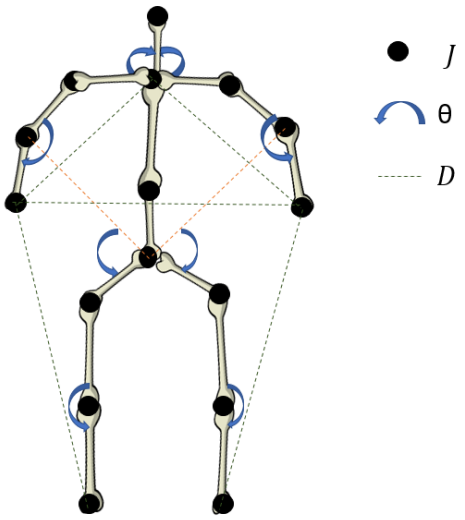


Figure 4.2: Representation of the selected characters for the body component.

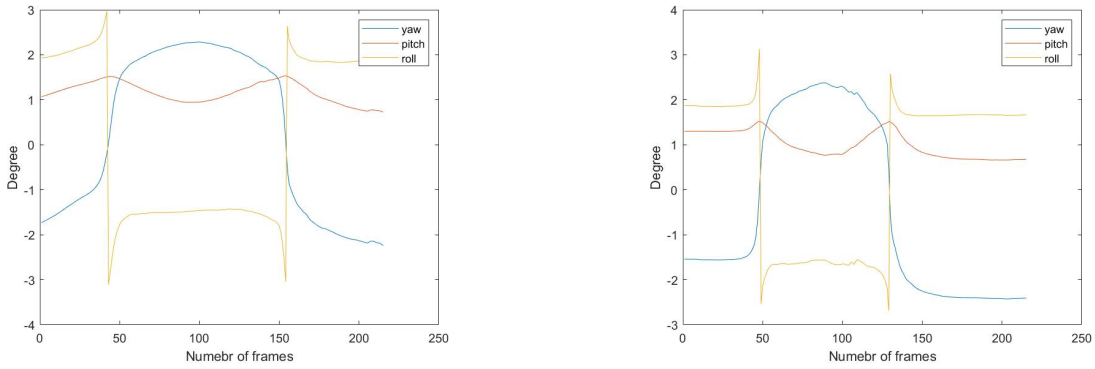


Figure 4.3: Changes in the yaw, pitch and roll of the hip center while sitting down.

Where $j_i = (x_i, y_i, z_i)$, $j_j = (x_j, y_j, z_j)$, and $j_k = (x_k, y_k, z_k)$ are the $3D$ coordinates of three consecutive joints. We also calculated the distances between the joints that move the most during a gesture according to the following formula.

$$d = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \tag{4.2}$$

Where i and j belong to the index of the joints of the head, left hand, right hand, left foot and right foot. Figure 4.2 shows all the selected elements for the body component.

The Space component describes the location, directions and paths of a movement. This component answers the question of where the body is going and in which space does it fit. In this part, we chose the changes of Yaw, Roll and Picth of the body as an element of the descriptor. With this choice, we can find out which axis the body rotates during a gesture. Figure 2 shows the changes in these three angles for the hip center during the "sitting down" gesture by two different people. As can be seen in this figure, the changes in these rotations have the same direction in two people.

The Shape component consists of three distinct qualities to describe the changes of the form of the body: shape flow, directional movement, and shaping. Shape flow reflects the relationship of the body with itself. The changes can be seen as the increasing or decreasing volume of the shape of the body. At this point, we've

selected a few geometric factors as follows:

- The area of $Tr = \{T_1, T_2, T_3, T_4, T_5, T_6, T_7\}$, in which Tr is the set of triangles formed by the three joints that have the most changes in different parts of body (Fig 4.4):
  - Triangle consisting of head, right foot and left foot, T_1.
  - Triangle consisting of hip center, right foot and left foot, T_2.
  - Triangle consisting of head,neck and right hand, T_3
  - Triangle consisting of head, neck and left hand, T_4.
  - Triangle consisting of right hand, left hand and hip center T_5.
  - Triangle consisting of left hand, left shoulder and left hip T_6.
  - Triangle consisting of right hand, right shoulder and right hip T_7.

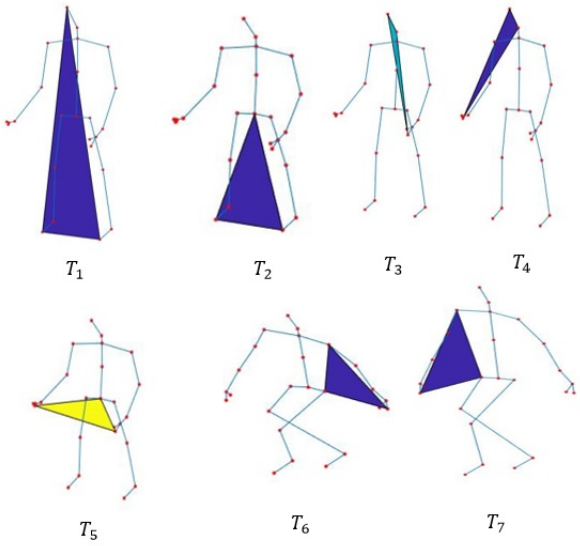


Figure 4.4: The defined triangles in the descriptor.

Shaping sub-component Characterizes the change in the shape of the body in relation to its space and defines the connection of where the body interacts continuously and three-dimensionally and the volume of the its surrounded environment. For this sub-component we calculate the volume of B, in wich B is the 3D boundary around all the joints of the skeleton data. In order to obtain the boundary that encapsulates all the joints of the skeleton body, we used the algorithm suggested in [257] (Fig 4.5). The last sub-component of shape is called Directional Movement which predict a where the body is directed toward some part of its environment. It is divided further into Straight-like and Arc-like trajectory. To do this end, we calculated the curvature measurement of the left and right hand according to the following formula:

$$C(j_h) = \frac{\|v(j_h) \times a(j_h)\|}{(\sqrt{v(j_h)_x^2 + v(j_h)_y^2 + v(j_h)_z^2})} \tag{4.3}$$

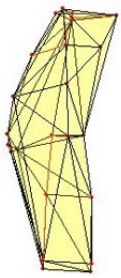


Figure 4.5: The 3D boundary around all the joints of the skeleton body.

Where, the h index indicates the left and right hand, the Vindicates the speed, and the a indicates the acceleration. If C tends to zero, it means that the path taken by the hands is a straight path, and a large curvature indicates a curvy trajectory.

4.2 Classification Via Deep Learning using LSTM

In order to introduce this step of our pipeline, we will first explain the Long Short Term Memory (LSTM) algorithm. LSTM is an advanced model of recurrent neural networks (RNN). These networks, as we will see in more details later, have a complex architecture but perform simple calculations, which essentially allows them to have two elements. On the one hand, they make time explicit through their architecture: they are considered to be close to the operation of Back Propagation Through Time (BPTT) in the way they manage sequential inputs. On the other hand, thanks to the "constant error Carrousel", a linear unit of calculation that we will present later, they can maintain the error over several time steps without it disappearing or being altered. This latter feature allows LSTMs to handle long runs in this way without suffering from the problem of vanishing gradient or any other unfolding problem. In this section, we describe this model in more details by presenting the architecture of an LSTM network, but also by detailing the structure of an LSTM unit. We will then discuss the calculations performed during forward propagation and the recommended algorithm for back propagation during training. Finally, in light of all of these technical details, we will explain the contribution of LSTMs to sequential learning. The concept of LSTM, for Long Short Term Memory, stems from the following idea: in an RNN, information is stored in two formats. The activation of the units represents the recent history of the network and therefore a short-term memory of the network, while the weight of the links between the neurons represents the experience accumulated by the RNN during its learning and therefore a long-term memory. Based on this observation, Hochreiter and Schmidhuber [96] introduced memory cells, a form of intermediate memory making it possible to store important information over a longer period of time than existing RNN [150]. From a technical standpoint, LSTMs are motivated by the desire to provide a model that does not suffer from the problem of overflow or fading out of learning error. In fact, in a traditional RNN such as Elman, the learning ends up inducing an error signal which either explodes (which induces increasing weights) or vanishes (especially when the network has to learn on long lead times which takes prohibitive time or does not work at all). Since 1997, research related to LSTM has been a very active field and many variations have been proposed. In the following section, we will present the version of LSTMs proposed in [84] and which is the first to have created a break in terms of performance with the original model.

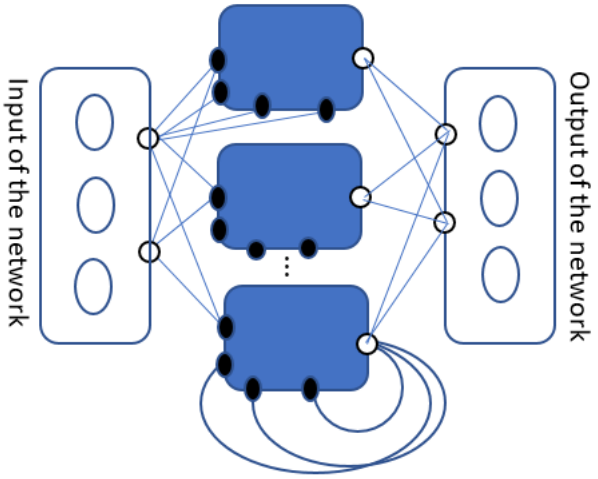


Figure 4.6: Presentation of a network with LSTM units at the hidden layer.

4.2.1 Architecture of an LSTM unit

How do LSTM units fit into an RNN-LSTM?

In an RNN such as Elman, the network consists of 3 layers: an input layer, a hidden layer and an output layer. The hidden layer is made up of artificial neurons, the same type as those in the input layer and those in the output layer. In an RNN-LSTM, the network architecture remains the same, namely 3 layers. The difference is basically in the hidden layer. The units used are then LSTM units made up of blocks and cells. The hidden layer thus receives at time t the information coming from the input layer, but also the information coming from the hidden layer at time $t-1$. The hidden layer activity pattern consists of the set of outputs from the different cells of each LSTM unit. Figure 4.6 shows an example of an LSTM.

The Blocks

Two concepts are important to distinguish when dealing with an LSTM: blocks and cells. If the first term designates the container, the one that learns to let in or out the information received or who chooses to forget it, the second term designates the content, that which processes the information. An LSTM block can have one or more cells: it distributes information to them equally. A cell, on the other hand, only belongs to one block. Figure 4.7 illustrates the case of a block with 2 cells and figure 4.8 that of a single cell block. An LSTM block, j, has the particularity of having $(3+c)$ data input channels, c denoting the number of cells in the block. The entry routes for information are the input gate in_j, the forget gate Φ_j and the output gate out_j, and the entrances to each cell. Each cell of each block receives the same information. We will detail this in the following section dedicated to the presentation of cells.

- The input gate: When its value is close to zero, this gate prevents external information from reaching the cells of the blocks and therefore from updating their value. Information from the lower layers is therefore not transmitted to the upper layers.

- The forget gate: When its value is close to zero, the cell forgets its past. The cells of the block then take

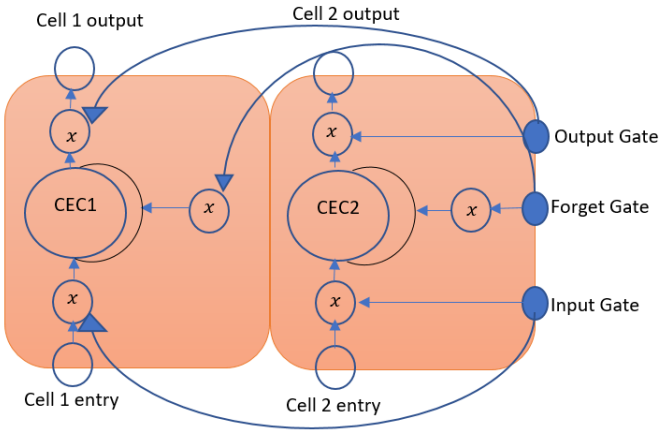


Figure 4.7: Presentation of an LSTM unit with a block composed of 2 cells.

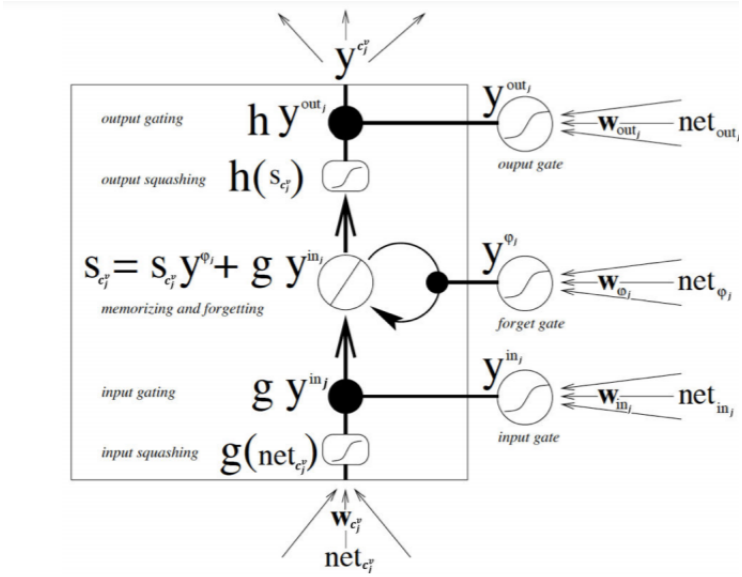


Figure 4.8: Presentation of an LSTM unit equipped with a block, itself equipped of a cell. Image taken from [84]

into account only the present item.

- The output gate: This gate determines whether a cell can send a value to other cells or not. Indeed, when its value is close to zero, the cell provides no output information (value also close to 0)

Each of the gates receives, at a time step t, the information associated with the time step t from the input layer, but also that associated with the time step t-1 from the cells of the same hidden layer. Although the learning algorithm is detailed below, it seems important to specify that the weights arriving on these gates are all modified during learning. This implies that each block learns to recognize when to take into account external information, but also when to communicate with the outside. Note also that the output gate described above were not part of the original model. They have been proposed by [84] and have since been proposed in the majority of implementations with a view to improving performance [150].

The cells A cell c, belonging to a block j and whose number is v, and which we will denote by c_j^v, also has an input channel and an output channel (in addition to the 3 channels linked to the 3 gates). In addition to this, each cell has a linear unit of calculation called "Constant Error Ccarousel", CEC, whose activation is noted $s_{c_j^v}$. This is called the state of the cell. It is this cell that solves the problem of vanishing gradient that prevents

learning. In the absence of new inputs or an error signal (during back propagation) reaching this cell, the local error feedback (induced by the recurring loop) of the CEC remains constant. It can therefore be maintained over a long period of time (we will come back to this later). Indeed, thanks to the gates of the block, the state of the cell is not updated at each time step. In terms of information received, a CEC cell takes into account different types of information: the information calculated at the time step t (the entry of the cell, that of the state of the cell $s_{c_j^v}(t)$, the values of the various gates $y^{in_j}(t)$, $y^{j}(t)$ and $y^{out_j}(t)$, but also information relating to the time step $t-1$ (the set of output values of all the cells of the layer hidden $y^{c_j^v}(t-1)$ and the state of cell $s_{c_j^v}(t-1)$. This two-step operation is what gives LSTM its performance.

4.2.2 Operation of an LSTM unit

An LSTM unit, as we have seen, is made up of several elements: 3 gates and as many inputs and outputs as cells. Despite the complexity of the architecture, performance is excellent thanks in particular to the simplicity of the calculations performed. Indeed, apart from the cell outputs, all of the other components (cell gate and cell entry) are artificial neurons, which use standard back propagation for their learning. And internally, calculations are products and sums, which makes everything easily differentiable and manipulable for learning. Another way to represent an LSTM unit is to think of it as a schematic representation described below (Algorithm 2). We will describe in this part the operation of an LSTM unit and the different equations needed for that.

Algorithm 2 Algorithm of how the gates work in the LSTM method

```
IF The value of input gate > 0 Then
    Open gate: Transmission of external information inside the block
        IF The value of forget gate > 0 Then
            The previous state of the CEC impacts the calculation of the present state
                IF The value of output gate > 0 Then
                    Open gate: Transmission of internal information to the outside of the unit
                ELSE
                        gate closed: No information provided on exit
                END
        ELSE
        The cell forgets its previous value
        END
ELSE
    gate closed: Information blocked
END
```

Notations and activation functions To start, it is necessary to define here some notations:

For blocks and cells:

- j: Block of a given hidden layer
- c: Cell in a given block
- c_j^v: Cell v of block j
- in_j: The input gate to block j
- Φ_j: The forget gate of block j
- out_j: The output gate of block j
- $s_{c_j^v}(t)$: The states of cell v of block j
- k: The Neuron at the output layer of the network
- e_k: The error calculated at the unit k level of the output layer of the network
- t^k: The output (value) expected at unit k

Other notations:

- m: The neuron having a link with an LSTM unit
- net_x: The activation function received by the calculation unit x (whether it is a gate or an artificial cell or neuron)
- y^m: The activity of computing unit m
- $w_{(l,m)}$: The weight coming from m and going to the calculation unit l

Activation functions:

- $\sigma : x \to \sigma(x)$: Sigmoid, logistic function between $[0, 1]$ such that

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{4.4}$$

- $g : x \to (x)$:

$$g(x) = \tanh(x) \tag{4.5}$$

4.2.2.1 Forward propagation

Reception by the network of information at time step t initiates the forward propagation of information in the network. This must then be transmitted from the input layer to the hidden layer and then to the output layer. At time t, each block of the hidden layer receives 2 pieces of information: the activity pattern of the input layer at time t, and the activity pattern of the hidden layer at time t-1 . In other words, the hidden layer returns to itself its own pattern of activity and this thanks to the information that each cell returns to itself, the other cells, and the doors of the different blocks. This previous and present data is received at the same time by

the input cell of the block, the input gate, the forget gate and the output gate. The calculations of activations received are detailed below:

At the inputs of the LSTM unit, block and cell: Let m be the number of neurons which have a link with an LSTM unit and which send information at time t, p the number of neurons which have a link with an LSTM unit and which send information at time $t-1$ and let be net_x the activations calculated for a calculation unit x: For the cell $x = c_j^v$:

$$net_{c_j^v} = \Sigma_m(w_{(c_j^v,m)}.y^m(t)) + \sigma_p(w_{(c_j^v,m).y^p(t-1)}) \tag{4.6}$$

For the input gate, $x = in$:

$$net_{in_j} = \Sigma_m(w_{(in_j,m)}.y^m(t)) + \sigma_p(w_{(in_j,m).y^p(t-1)}) \tag{4.7}$$

For the forget gate, $x = \Phi$:

$$net_{\Phi_j} = \Sigma_m(w_{(\Phi_j,m)}.y^m(t)) + \sigma_p(w_{(\Phi_j,m).y^p(t-1)}) \tag{4.8}$$

For the forget gate, $x = out$:

$$net_{out_j} = \Sigma_m(w_{(out_j,m)}.y^m(t)) + \sigma_p(w_{(out_j,m).y^p(t-1)}) \tag{4.9}$$

For each of the gates, the activation function σ is applied to the inputs received such as:

$$y^{\Phi_j}(t) = \sigma_{\Phi_j}(net_{\Phi_j}(t)) \tag{4.10}$$

$$y^{\Phi_j}(t) = \sigma_{\Phi_j}(net_{\Phi_j}(t)) \tag{4.11}$$

$$y^{out_j}(t) = \sigma_{out_j}(net_{out_j}(t)) \tag{4.12}$$

And

$$y^{c_j^v}(t) = \sigma_{c_j^v}(net_{c_j^v}(t)) \tag{4.13}$$

In the LSTM unit, at cell level: Updating the cell status can be described in the sequence of following calculation:

- Step 1: The activity of the input of cell $g(net_{c_j^v})$ is multiplied by the value of the input gate of block $y_{in_j}(t)$. This is where the block can prevent irrelevant information from changing the internal state of the cell. When the front gate is closed (close to zero activation), irrelevant inputs, as well as noise, do not enter the LSTM unit.

$$y^{c_j^v}(t)y^{in_j}(t) = g(net_{c_j^v}(t)).y^{in_j}(t) \tag{4.14}$$

- Step 2: The forget gate acts as a reset button for the past. For example, when the study of a new sequence

begins, its activity is close to zero. Since the product of Equation 4.13 is then zero, the cell "forgets its past" and can begin to store a new past.

- Step 3: This is the sum of the two previous steps: in other words, taking into account the present knowing the past (or the absence of the past when it is the beginning of a sequence).

  The equation for updating the internal state of the cell at time t is therefore as follows:

$$S_{c_j^v}(t) = g(net_{c_j^v}(t)).y^{in_j}(t) + S_{c_j^v}(t-1).y^{\Phi_j}(t) \tag{4.15}$$

  This equation 4.15, carried out for each cell, made it possible to save calculation time. This equation is only performed if all the gates have values greater than 0. If the first two conditions (input gate and forget gate) are not met, the CEC does not receive entry. Not being updated, the information within it is preserved for this time step without degradation.

- Step 4: Once the internal state has been calculated, all that remains is to calculate the output of the cell, which involves the output gate of the block as follows:

$$y^{c_j^v}(t) = \sigma(S_{c_j^v}(t)).y^{out_j}(t) \tag{4.16}$$

4.2.2.2 Back propagation

In an RNN-LSTM, the output units use BPTT for the backward propagation of the gradient. Block output gates use a truncated version of BPTT and finally the weights going to block cells, entry gates and forget gates will be updated via a truncated version of Real-time recurrent learning (RTRL) [277].This choice is justified by the fact that thanks to the truncation, all errors will be cut when they exit a cell or gate, although they are used to modify incoming weights. Thus a cell's CEC becomes the only part of the system through which the error can be back-propagated and maintained forever. This makes LSTM updates efficient without significantly affecting the learning power: the error flow out of cells tends to decrease exponentially anyway [97].

Calculation of derivatives during forward propagation: Back propagation calculations begin during forward propagation. Indeed, at each time step, after the calculation of the output of the network according to an input at time t, the RNN-LSTM model also calculates partial derivatives to calculate the gradient and that for the input gates, the forget gates and cells of each block. The idea is to calculate the derivative of the value of the cells relative to the network weights for these information entry channels there. So for each block j, the derivatives are calculated as follows: For cells:

$$dS_{c,m}^{jv}(t) = dS_{c,m}^{jv}(t-1) * y^{\phi_j}(t) + f'(net_{c_j^v})(t) * y^{in_j}(t) * y^m(t-1) \tag{4.17}$$

For the input gates:

$$dS^{jv}_{in,m}(t) = dS^{jv}_{in,m}(t-1) * y^{\phi_j}(t) + g(net_{c^v_j})(t) * f'_{in_j}(t) * y^m(t-1) \tag{4.18}$$

And for forget gates:

$$dS^{jv}_{\Phi,m}(t) = dS^{jv}_{\Phi,m}(t-1) * y^{\phi_j}(t) + g(S_{c^v_j})(t) * f'_{\Phi_j}(net_{\Phi_j})(t) * y^m(t-1) \tag{4.19}$$

In which at time $t = 0$, $dS^{jv}_{c,m}(0) = 0$, $dS^{jv}_{in,m}(0) = 0$ and $dS^{jv}_{\Phi,m(0)=0\}}$.

Using the proposed descriptor as well as the LSTM algorithm , our proposed framework is built as shown in Figure 4.9. The input of the first block of LSTM are the spatio-temporal feature described in previous section. In order to prove the efficiency of the proposed framework, in the last step, a softmax layer is placed on the last layer of the network to approximate the probability that the sequence of X belongs to class C_k:

$$P(C_k|X) = \frac{\exp(z_k)}{\Sigma_{i=1}^{C} \exp(z_i)} \tag{4.20}$$

Where

$$z = W_z h_n + b_z \tag{4.21}$$

Where h_n is the output of the fully connected layer from LSTM structure. In order to evaluate the proposed

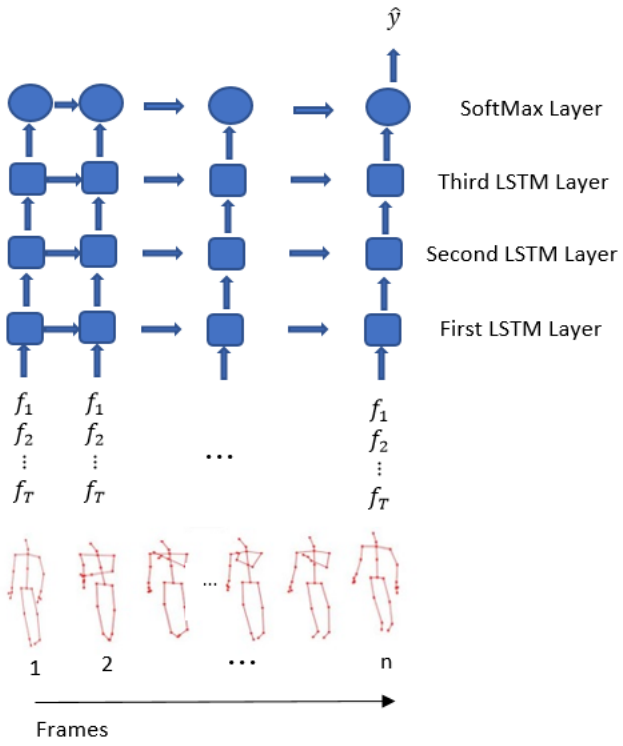


Figure 4.9: Demonstration of The Proposed Pipeline.

framework in this chapter, two datasets, NTURGB+D 120 [151] and SYSU HOI [102], were examined which will be introduced in the following.

4.3 Experiments

Since the selected factors in the proposed descriptor are invariant to the coordinate system, there is no need to normalize them. And these type of features reduce the complexity of the calculations. In this work, an LSTM algorithm is implemented in order to train data. The learning rate and momentum are set 0.01 and 0.9 respectively [240]. In the following, we demonstrate the experiments of implementing the proposed gesture recognition algorithm based on spatio-temporal features. In the following, we will briefly introduce the evaluated data sets and then compare the results obtained from the implementation of the proposed algorithm on these two data sets with other works in the literature.

4.3.1 Implementation Details

In our experiments, each video sequence is divided to T sub-sequences with the same length, and one frame is randomly selected from each sub-sequence.

The implementation of our approach is based on the MATLAB 2018. In both approaches, we implemented our proposed descriptor before the first LSTM block of the designed network. For the both approach we used "Holdout" validation with a fraction of 0.3 for SYSU HOI, MSRACTION, Florence3D, UTKinect and 0.4 for NTU 120. We set the number of hidden layer to 50 for SYSU HOI, MSRACTION, Florence3D, UTKinect and 100 for NTU 120. For all data set the learning rate is set to 0.01. The hardware used in this work is a single GPU.

4.3.2 Results

In this section, we compare our approach with recent researches. The obtained accuracy of recognition on SYSU HOI and NTU-120 datasets are reported in Table 4.1. Figure 4.10 also shows the Confusion matrix for the SYSU HOI data set.

Method	SYSU HOI	NTU 120
Cooperative Training of Deep Aggregation Networks [266]	98.33	-
Deep Bilinear[103]	88.9	-
Dynamic Time Warping [199]	92.32	-
Convolutional Neural Networks with Joint Supervision [147]	97.08	-
GVFE + ST-GCN w/ DH-TCN[182]	-	**74.2**
Body Pose Evolution Map[152]	-	66.9
Skeleton Visualization (Single Stream)[153]	-	63.2
Multi-Task Learning Network[118]	-	57.9
Our Method	**98.94**	72.6

Table 4.1: Accuracy of recognition (%) on SYSU HOI and NTU-120 datasets. The evaluation is performed using holdout validation.

In the following diagram 4.11, the differences between the results obtained from the implementation of the method presented in this section and the module of the previous section are compared. As can be seen in all 4 data-sets, the module presented in this section has better results. The following are the configuration matrices of MSRACTION, FLORENCE3D and UTKinect data-sets, respectively.

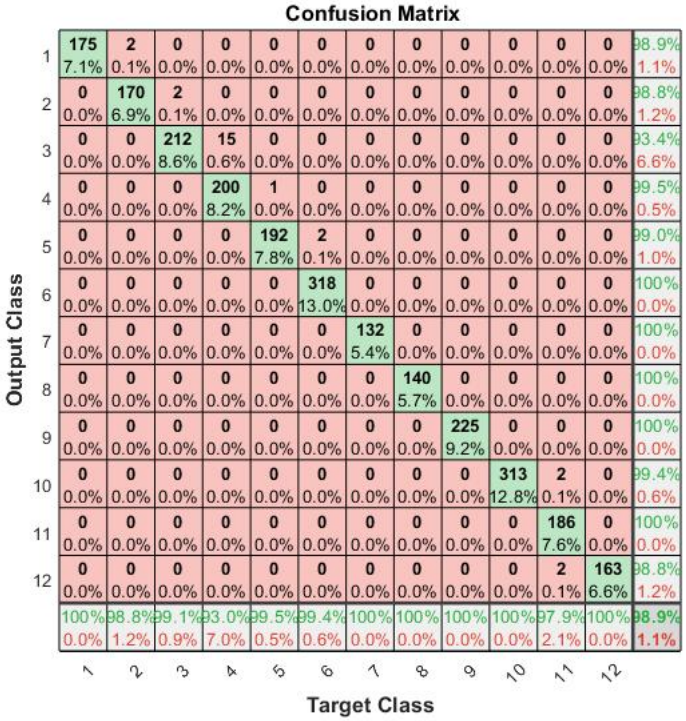


Figure 4.10: Confusion matrix of SYSU HOI data-set.

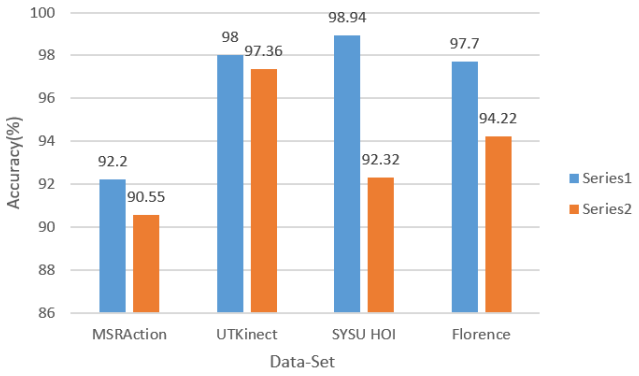


Figure 4.11: Comparison of results obtained by LSTM (Series 1) method and dynamic time warping (Series 2).

4.4 Conclusion

In this chapter, we focused on recognizing human movements using geometric and time-dependent features. Deep learning methods basically accept raw data and train them in designed networks. Because the data sets available for use in deep learning methods are large, performing processes such as normalizing 3D coordinate data is time consuming and complex. Raw data that is not normalized to the initial location and angle of the person performing the gesture requires more training, more neurons, more layers, as well as more complex calculations in deep learning training networks. For this reason, we decided to select components as descriptive components that are independent of the original location of the person performing the gesture. In this regard, in addition to the distances between hip center and other joints of the body and the changes of the quaternion angles in time, we define the triangles formed by the different parts of the body and calculated their areas. We also calculated the area of the single conforming 3-D boundary around all the joints of the body. At the end

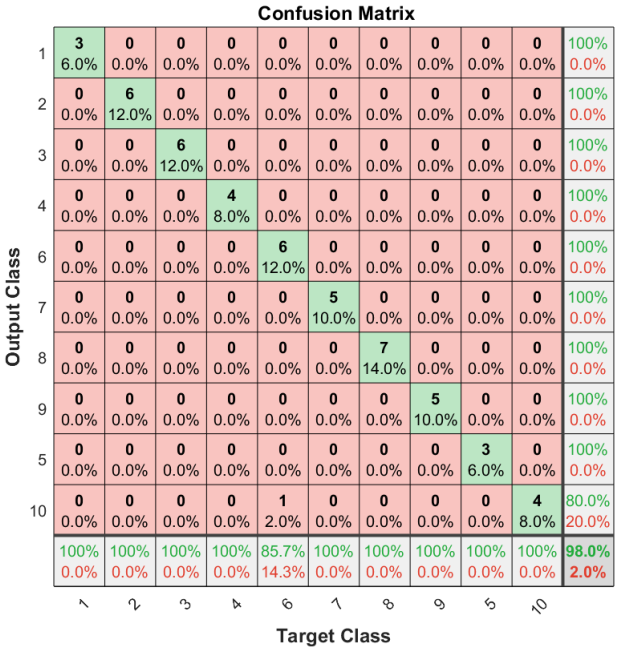


Figure 4.12: Confusion matrix of UTkinect data-set.

we added the velocity of different joint in the proposed descriptor. By doing this, in addition to increasing the speed of data training, we also reduce the error rate due to non-normalization of data. We used the LSTM method to train this data. In order to evaluate the performance of the proposed pipeline, we implemented it on two data sets. One of the aims of this thesis is to create a bi lateral interaction between humans and robots, and the proposed methods should correspond to the natural conditions of life, so we decided to test this method on a small data set to ensure its effectiveness. Since the SYSU HOI data set consists of daily movements, it has been selected as a small data set. The NTURGB+D 120 data set has been selected as a large data set to evaluate the proposed method. The results obtained are compared with the results of similar work on these two datasets, and as can be seen, in many cases our proposed method has a better result.

In the next chapter, by expanding the descriptor and also making the pipeline more developed, we will recognize the emotions of human according to their body movements.

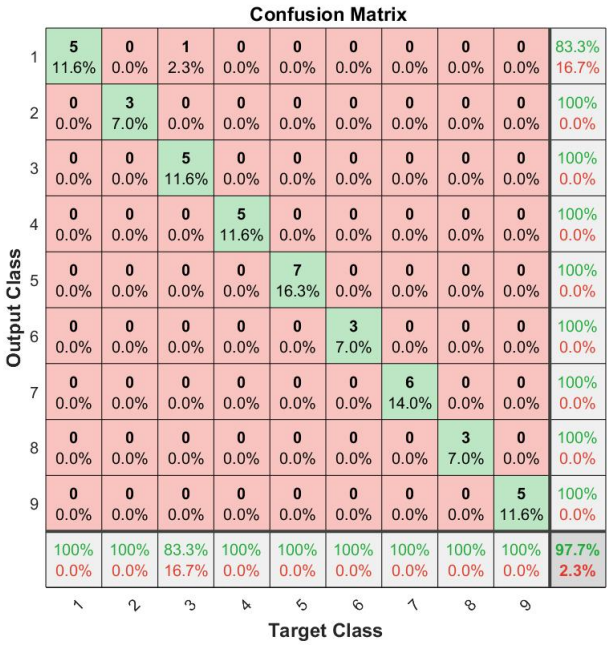


Figure 4.13: Confusion matrix of FLORENCE 3D data-set.

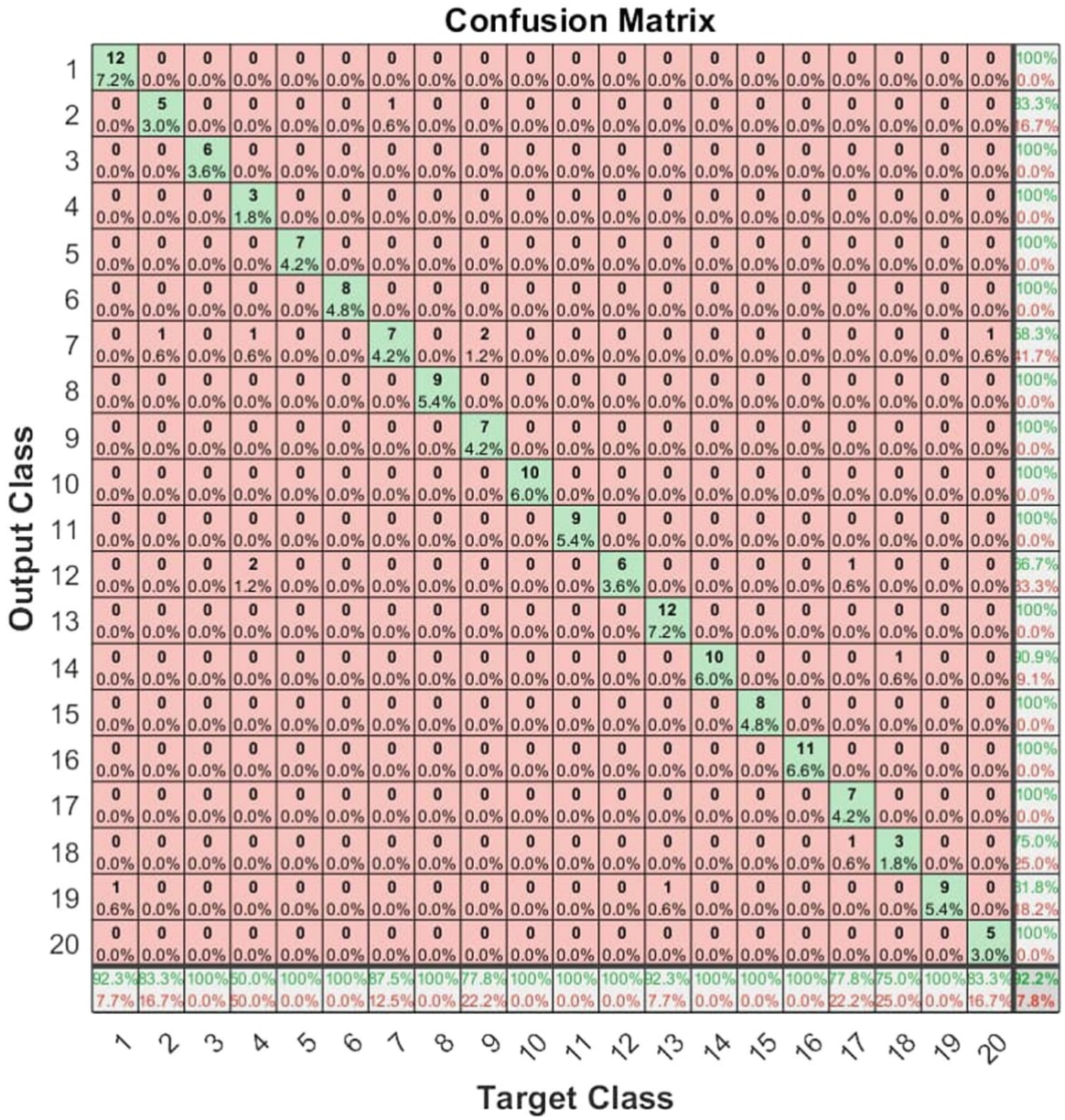


Figure 4.14: Confusion matrix of MSRAction 3D data-set.

Chapter 5

Recognition of expressive gestures

Contents

In this chapter, we develop a system for recognizing expressive gestures. We integrate here the term of expressiveness to improve our robotic application "Human-Robot Interaction" and make it more natural. Our system will have to be able to recognize the gesture of the person and also his emotional state through his movement. We are building a database with the MVN Awinda motion sensor from Xsens, consisting of 5 gestures interpreted with 4 different emotions (joy, anger, sadness and neutral). This chapter is done in two separate parts. The descriptor used in the first part is made using the LMA method. Random Decision Forest (RDF) has also been used to identify and classify emotions. In the second part of this chapter, the descriptor is based on the geometric coordinates of the body skeleton and time-dependent characters and the recognition part is based on Feed Forward Neural Network.

As we said, the descriptor is made using LMA. Since the purpose of this section is to identify emotions, in addition to the three factors, body, shape and space, the Effort factor has also been added in this section. For the training and classification phases, we use the machine learning library (Scikit-learn). We choose four of the most famous learning methods (decision tree forests, multilayer perceptron, support vector machines: One-Against-One and One-Against-All). An adjustment of the various parameters of the models is carried out in order to have a better performance of the system. A comparison between the different classification algorithms is made in order to choose the best one. We evaluate our approach with some public data set, as well as our expressive gesture base. This section of this chapter is organized as follows: first sub section presents the setup of the expressive gestures database. In the next, we expose our global motion descriptor inspired by the LMA method. Next we concern the part of classification of expressive gestures by the use of Random Decision Forest (RDF). We conclude in final sub section.

5.1 Construction of the expressive gestures database

5.1.1 Description of the Xsens Expressive Motions (XEM) data set.

Our XEM (5 Expressive Control Motions) data set consists of five expressive movements (dancing, walking, waving, pointing and stopping), shown in Figure 5.1. The purpose of selecting these gestures was to evaluate the recognition of emotions based on body movements. For this reason, the gestures are chosen to cover different situations. As will be seen in the results, the 3 senses we face daily can be understood through how to perform a hand-only gesture (pointing). We captured neutral movements, as well as movements made with emotions: joy, anger and sadness. Each expressive gesture is repeated 5 times.

5.1.1.1 The participants

11 people (five men and six women) from the University of Evry Val d'Essonne, aged 27 to 36 (average = 29.85 years, standard deviation = 2.47) participated in the construction of our data set. The Motion capture sessions were recorded with permission from the participants. Our corpus is well anonymized, collected with the consent of the various participants. All actors received monetary compensation for their participation, and were

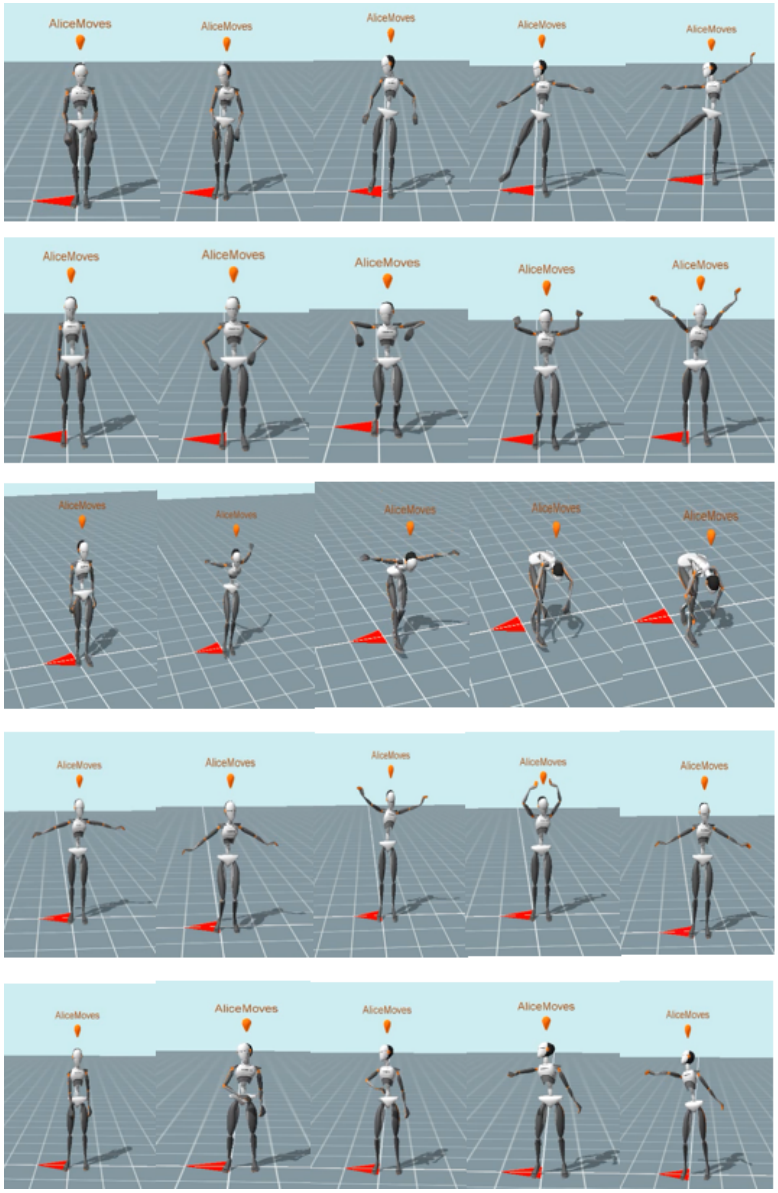


Figure 5.1: The XEM data set, the top-down gestures are: dance, move forward, stop, wave and point.

informed that the data collected would be anonymized, shared and analyzed strictly for research purposes. In order to avoid any exaggeration of the expression of emotions, we have chosen people who are not professional actors, which allows to have an application adaptable to everyone.

5.1.1.2 Scenarios

To help the participants to express their emotions well, we offer scenarios presenting emotional situations and we played music. Each participant was invited to read the proposed scenario, take the time to soak up the situation and, as soon as they feel ready, perform the requested action 5 times. Examples of proposed scenarios for the emotions of joy, sadness and anger are shown below:

- Emotion of joy
  - You've passed all of your college courses and the same day you got your results, you get a phone call from a company for a job with a really good salary.
  - You are in a good relationship. Your partner's birthday is near and you are planning to have a party, but you don't have enough money, suddenly you get an email that informs you that you have won in a lottery.
- Emotion of sadness
  - For a popular festival, you have booked your ticket and hotel 6 months in advance, but due to a sudden event, you miss your flight and therefore can not reach your favorite festival.
  - You have lost someone very close to you in an accident.
- Emotion of anger
  - You have a job interview you've been waiting for a long time. Today is daylight saving time and you forget to change your clock. So you wake up late, get dressed quickly and get in your car that you have found broken down, so you have no other choice but public transport. Finally, in the metro station, you find that the drivers are on strike.
  - You have your own restaurant that loses more money than it earns. You decide to verify the cause, then you introduce yourself as a new hire to your employees. In the first week you notice that the restaurant is open at a late hour and therefore there is always a long lead time for customer orders. You discover unpleasant service and behavior from servers towards customers. At the end of the day, you are very angry with the ineffectiveness of your employees.

During the recording sessions, the order of scenarios, emotions and movements was randomized from one participant to another.

Figure 5.2: The MVN Awinda sensor from Xsens.

5.1.1.3 Materials

For the acquisition of expressive gestures, we used a camera-free "motion capture" product supplied by the Xsens group, which is the MVN Awinda system [204]. This is a combination with sensors attached by straps for tracking 17 joints (head, sternum, pelvis, left / right shoulders, left / right forearm, left / right rear arm, left / right hands / right, front left / right legs, rear left / right legs, left / right feet). This system is made up of 17 inertial units (Fig 5.2). Using this 17 sensors as well as the distance between different parts of the body that must be calculated when installing sensors on the body of the participants, Xsens gives us the specifications of 23 different body segments (estimated data).

Each inertial unit is made up of an accelerometer, a gyroscope, magnetometer and barometer. The recent addition of GNSS (Global Navigation Satellite System) to Xsens MVN allows position aiding to inertial motion capture. Position and orientation information from the sensors in real time, at 60 frames per second, is recorded. In addition, this system combines two algorithms, the Kalman XKF3-hm filter and the SDI (Strap-Down Integration) algorithm, which provide precise information on the positions and orientations of the joints. This system has two major advantages over the Kinect sensor. It does not require any particular lighting condition for its operation. In addition, with this system there are no limitations that can be obtained by occlusion with surrounding objects or people interacting with the actor. In addition, by putting on the suit without a camera, actors are not constrained to a specific measurement volume and their movements can be measured in a familiar environment while performing their tasks, such as in everyday life.

5.1.2 Format of motion capture files

For each participant, we collected MVNX (Moven Open XML format) files. It is an XML format that can be imported to other software, including MATLAB and Excel. This format contains several information, including sensor data, segment kinematics, and joint angles, as well as subject information needed to recreate a 3D visualization of an avatar.

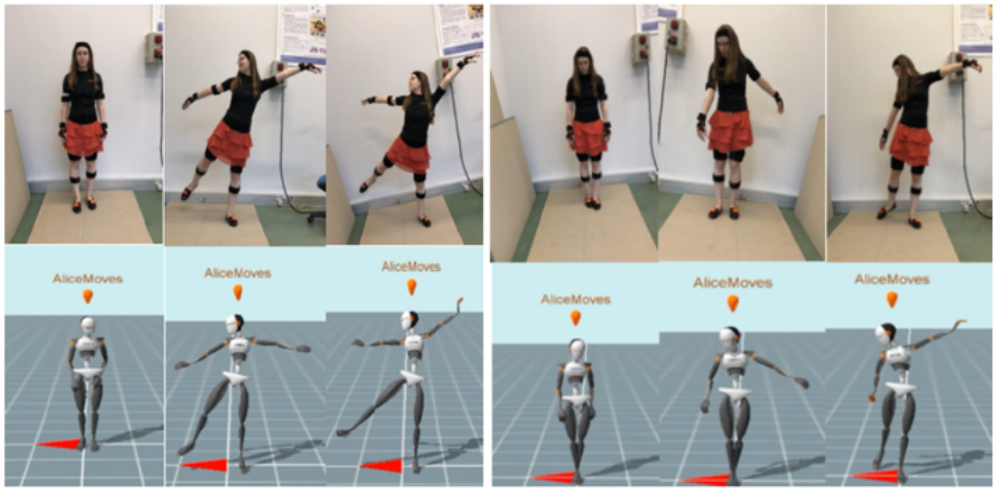


Figure 5.3: The dance gesture performed with two different emotions: a) Happiness b) Sadness.

5.2 Expressive global descriptor inspired by LMA

5.2.1 Effort-Shape relationship

Laban has developed an interesting model for describing and analyzing dynamics and styles of movement, called the Effort-Forme model [87]. This model reflects the internal state of the person during the execution of the movement by relating a set of physical properties of the movement with expressive qualities, such as weight, time, etc [25]. It has been adopted in several contexts, notably for the analysis of expressive movements in dance [25, 12, 10, 88, 90, 15, 46, 45], in music [46, 39, 45] and in medicine [74], etc. Moreover, this model has sparked an interest in the computational modeling of virtual agents for a more realistic and expressive animation [53, 116]. Effort concerns the observable dynamic rhythms of physical effort and the phraseology of body movement [157]. Effort reflects the inner attitude towards the use of energy according to its four factors: space, time, weight and flow. These 4 factors, when arranged in a specific way, create the eight basic Effort actions (See Table 5.1).

	Space	Time	Weight	Flow
Punch	Direct	Suddenly	Strong	Bound
Press	Direct	sustained	Strong	Bound
Cut	Indirect	Suddenly	Strong	Bound
Twist	Indirect	sustained	Strong	Bound
Touch	Direct	Suddenly	Lightweight	Free
Slide	Direct	sustained	Lightweight	Free
Browse	Indirect	Suddenly	Lightweight	Free
Float	Indirect	Sustained	Lightweight	Free

Table 5.1: The eight elementary actions of E ort.

Each factor being a continuum between two elements of Effort, either space (indirect / direct), time (sudden / sustained), weight (light / strong), or Flow (free / bound). The Effort/Shape relationship allows attention to be focused on two aspects of bodily movements: on the one hand, how kinetic energy is spent in space, and

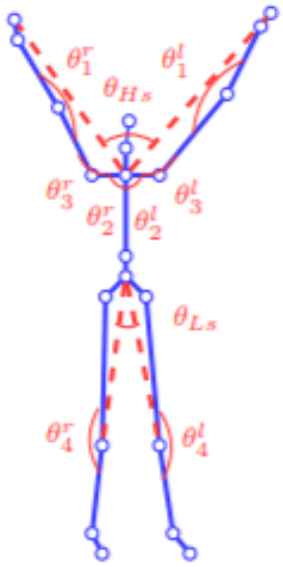


Figure 5.4: Body characteristics.

force and time in functional and expressive behavior. On the other hand, the shape of a movement, or how the body changes and moves in space.

5.2.2 Global descriptor

In order to describe the whole gesture, we use global measures based on the components of LMA presented in chapter chapter 3. We quantify the Effort component to capture the qualitative and expressive aspect of the gesture. We start with the component of the Body which has the responsibility of highlighting the body part which is moving, making the connection between the moving parts and taking into consideration the issues of locomotion and kinematics. For this category, we describe the organization and connection between the different joints (see Figure 5.4). We consider two parts, the upper and lower part. For the first one, the extension of the different joints is described by computing the following angles in the left and right parts respectively: between hands and shoulders (θ_1^l, θ_1^r), between elbows and hips (θ_2^l, θ_2^r), between elbows and shoulders in the symmetrical part (θ_3^l, θ_3^r). We also calculate the distances between two hands (d_{Hs}) as well as the distances between the shoulder center and both hands $(d_{shc,lh}, d_{shc,rh})$. For the lower part of the body, the extension of the knees has been described with the angles between feet and hips (θ_4^l, θ_4^r). These two characteristics allow to characterize specific actions like crouch or hide gestures. We also characterize the opening of the legs with the angle computed between two knees θ_{Ls}. Space component describes the trajectory performed by the participant's body parts during an action. We compute the length (L) of the trajectory made by the upper extremities including head, left and right hands joints (J_t is the $3D$ joint position captured at frame t, and T is the number of frames in a motion sequence.

$$L = \Sigma_{t=1}^{T-1} \|J_{t+1} - J_t\| \tag{5.1}$$

Shape component can be defined as a collection of the properties that appear from body and space components. This category expresses the changing form of the body that is either motivated by self or the environment. The three factors of Shape are Shape Flow, which represents a relationship of the body to itself; Directional movement, which defines a relationship of the body toward some part of the surrounding environment; and

Shaping, which describes the qualitative changes in the shape according to three-dimensional planes. For the Shape Flow, we compute the convex hull volume (V) of the $3D$ skeleton based on Quickhull algorithm [22] in order to describe body extension. To form the descriptor, we calculate the mean (M_c), the standard deviation (S_c) and the range (R_c) of the elements listed so far as follow:

$$M_c^k = \frac{1}{T}\Sigma_{t=1}^{T} c_t^k \tag{5.2}$$

$$S_c^k = \sqrt{\frac{1}{T}\Sigma_{t=1}^{T}(c_t^k - M_c^k)^2} \tag{5.3}$$

$$R_c^k = \min_{1\leq t\leq T} c_t^k - \max_{1\leq t\leq T} c_t^k \tag{5.4}$$

c_t^k corresponds to the local characteristic defined so far. $k \in \{1,\ldots,l\}$ calculated at time t and l is the number of characteristics. For Directional factor, we focus on the upper extremities including, head, left and right hands. We describe their pathway by computing the gradual angular change ϕ occurring between two successive frames, defined as follow:

$$\phi_{J_t} = \arccos\left(\frac{\overrightarrow{J_{t-1}J_t}}{\|J_{t-1}J_t\|}.\frac{\overrightarrow{J_tJ_{t+1}}}{\|J_tJ_{t+1}\|}\right) \tag{5.5}$$

This equation describes the local curvature of joints trajectory. From this angle we derived the curvature feature (C) defined as:

$$C = \Sigma_{t=2}^{T-1}\phi_{J_t} \tag{5.6}$$

This index tends to 0 in straight-Line trajectories cases and changes to a very high value in curved paths. For Shaping factor, we describe movements according to three planes, Horizontal (sideways movements), Sagittal (forward/backward movements), and Frontal (upward/downward movements). We describe body inflation according to the three planes by computing average distances of all skeleton joints $[J_{t,i}]_{X,Y,Z}(t = 1,\ldots,T and i = 1,\ldots,N)$ with respect to the spine joint ($J_{1,s}$) at initial frame.

$$D_H = \frac{1}{T}\Sigma_{t=1}^{T}\left(\sqrt{\Sigma_{i=1}^{N}([J_{t,i}]_X - [J_{1,s}]_X)^2}\right. \tag{5.7}$$

$$D_F = \frac{1}{T}\Sigma_{t=1}^{T}\left(\sqrt{\Sigma_{i=1}^{N}([J_{t,i}]_Y - [J_{1,s}]_Y)^2}\right. \tag{5.8}$$

$$D_S = \frac{1}{T}\Sigma_{t=1}^{T}\left(\sqrt{\Sigma_{i=1}^{N}([J_{t,i}]_Z - [J_{1,s}]_Z)^2}\right. \tag{5.9}$$

The intention and dynamic qualities of the movement, the feeling tone, the texture and the manner of consumed energy during movement are explained by the Effort component. This component is generally associated with the change in emotion or mood, hence, it is useful for motion expressivity. This factor is required for the description of emotions. We focus on the upper body part since this is the part that moves the most when expressing an emotion, especially the following four joints: head, spine, left and right hands. The Effort component consists of

Figure 5.5: The factors of the Effort component (Space, Time, Weight and Flow).

four factors: Space, Weight, Time, and Flow which are explained as follows: Space expresses the quality of active attention to the surroundings. It is divided into two categories, Direct (when the action is direct the attention is on a single point in space, focused and specific) and Indirect (giving attention to multiple directions in the space, multi-focused and flexible). We compute the Straightness index (S) of upper body joints trajectories (Sh for head, Ss for spine, Sl for left hand, and Sr for right hand) as the proportion between the Euclidean distance of the straight trajectory between the positions at the first and last frame (D), and the sum of distances between successive frames (L) defined in the Equation 5.1. The Effort is the quality, the emotions and the inner attitude which is expressed by the movement. It is often described as the dynamics of movement, the qualitative use of energy. For example, if we look at the two actions "pushing a heavy object" and "closing a door", the two are very similar in terms of body organization. Indeed, the two actions are based on the extension of the arm. On the other hand, the attentions paid to the force of the movement, the control of the movement and the duration of the movement are very different. Laban proposed that the dynamics of human movement be summarized by a combination of the following factors, each of which has two opposite polarities (Figure 5.5):

- Space (Indirect/Direct): the "where" of the movement; attention / thought.
- Time (sustained / sudden): the "when" of the movement; intuition.
- Weight (light / strong): the "what" of movement; sensation.
- Flow (free / linked): the "how" of movement; feeling.

In our case, to characterize expressive gestures, we have assumed that the upper part of the body is the most expressive and most moving part in a control application. So, we focused more on the upper part specifically the following 4 joints: the head, the right hand, the left hand and the torso.Space expresses the quality of the attention that the person pays to the surroundings, and differentiates a Direct movement (when the action is direct, the attention is focused on a single point in space, targeted and specific) or Indirect movement (paying attention to multiple directions in space, multi-focused and flexible). The authors in [159, 160], quantified this factor with the calculation of the direction of the head. Likewise the authors in [14], considered that the movement is direct if the skeleton moves in the same direction as the orientation of the head, otherwise it is classified as indirect. For this they characterized the Space factor by measuring the orientation of the head, by calculating the angle between the orientation of the head and the trajectory of the artist's body which is defined by the trajectory of the joint from the center of the hip. We characterize this quality with the index

(S^k) of the straightness of the trajectories of the joints (k) of the upper body. This index is expressed by the ratio between the Euclidean distance of the two positions of the first and the last frame (D^k) and the sum of the distances between successive frames (L^k), calculated with the following equation:

$$S^k = \frac{D^k}{L^k} = \frac{\left\| J_T^k - J_1^k \right\|}{\Sigma_{t=1}^{T-1} \left\| J_{t+1}^k - J_1^k \right\|} \tag{5.10}$$

In a direct (rectilinear) movement, we get a straightness index close to 1 and in the case of an indirect movement, the value of S^k will be close to 0. Time describes the rhythm of the movement relative to its urgency and therefore distinguishes a Sudden movement (rapid, urgent, unexpected, surprising) from a Sustained movement (stable, continuous). In [211], the time factor is determined by the accelerations accumulated over time in the parts of the body. In [159, 160], the authors characterized the time factor by the angular speed of all the joints. The authors in [251], used 8 characteristics to quantify the time, the total duration of the gesture, the percentage of the pause periods relative to the entire sequence of the gesture. In addition, they took the two series which correspond respectively to the duration of the breaks and to the duration of the periods of activity and calculated for each the following three parameters: the mean, the standard deviation and the maximum value. The authors in [116], assumed that sudden movement is characterized by peaks of acceleration and movement sustained by a uniform speed (no acceleration). So to quantify the time factor, they measured the net acceleration accumulated over the duration of an interval of motion. The authors in [14], introduced 5 characteristics which correspond to the speeds and accelerations of the joints of the hands, feet and pelvis. In our case, to quantify this factor we compute the mean, the standard deviation, and the range of velocity (v) of upper body joints (head (v_h), spine (v_s), left and right hands (v_l, v_r)). Weight involves the strength or power with which a movement is performed. It can be strong or light. Strong movement requires force and acceleration while light movement is characterized by an invariant rhythm of motion, so lower acceleration values. We compute the mean, standard deviation, and the range of the acceleration (a) of the head (a_h), spine (a_s) and hands (a_l, a_r) to quantify the Weight factor. Flow involves the continuity of the movement. Its categories are bound (controlled movement) and free (unstoppable, liberated). We compute the yaw and $pitch$ ranges of the head ($rangeyaw_h$,$rangepitch_h$), spine ($rangeyaw_s$,$rangepitch_s$), left hand ($rangeyaw_l$,$rangepitch_l$), and right hand (range yaw_r,$rangepitch_r$) motions. Bound movement can be described as a movement which can be stopped at any moment, whereas free or smooth movement presents a continuity in the motion. Once we have quantified all LMA factors, we obtain a descriptor vector composed of 85 features which will be the input of the classification method presented in the next section.

5.3 Expressive motion recognition and analysis using the learning method

5.3.1 Recognition of expressive gestures with the RDF method

Generally, small changes in the learning base can lead to large changes in the construction of the classifier. This therefore requires a combination of complementary classifiers. The RDF method is based on the principle of building sets of diverse classifiers by a combination rule. This explains its superiority over other methods in

Figure 5.6: Statistical motivation.

the previous experimental parts. A lot of research is being done to study the advantages and the motivations in the use of this approach, called "ensemble of classifiers". Combining several classifiers is the same as using several classifiers and combining them in order to obtain a classifier which surpasses each of them. This type of approach is intuitive since it mimics our way of seeking several opinions before making a crucial decision [205]. The research area of "multiple classification systems" (MCS) has become very popular in the field of machine learning. Several articles on the construction of sets of classifiers have been published [65, 80, 253] and have shown the efficiency of this type of model to improve the generalization capacity of a single classifier, and thus reduce variance and improve precision. Improving performance in multiple classification systems is based on the concept of diversity, which states that a good set is one in which misclassified examples are different from one individual classifier to another. Therefore, various strategies are used to obtain a group of classifiers based on diversity. This diversity is achieved in several ways, for example, sub-resampling of training data, selection of subsets of objects, etc. According to [80], there are three main motivations for combining classifiers:

- Statistical reason: a learning algorithm can be seen as a search in a space H of classifiers to identify the best classifier. The statistical problem arises when the training data set is too small compared to the size of the classifier space. In this case, without sufficient data, the learning algorithm can find many different classifiers in H which all give the same precision on the training data. However, they can have different generalization performance. This can therefore increase the risk of selecting a bad classifier, i.e. a classifier with a bad generalization capacity. So the solution is to build a set of all these high-performance classifiers by combining their outputs. Dietterich [65] gives an illustration of this situation in Figure 5.6. The outer curve designates the space of classifiers H. The internal curve designates the set of classifiers which are efficient in learning. The point marked f presents the real classifier. The combination of precise classifiers makes it possible to give a good approximation of f.

- Reason for Computation: several learning models are based on local optimization techniques, which makes the model sensitive to local optima. Dietterich cited two examples, the neural networks algorithm that uses gradient descent to minimize an error function on training data, and decision tree algorithms that employ a fractionation rule in a way recursively to expand the decision tree. In the event that there is enough training data (so that the statistical problem is absent), it can still be very computationally difficult for the training algorithm to find the best hypothesis. A set built by performing the local search from many different starting points can produce a better approximation of the true unknown function, as shown in Figure 5.7.

Figure 5.7: Calculation motivation.

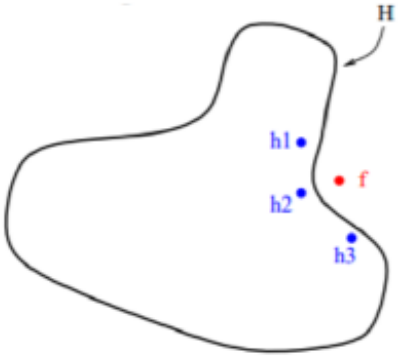


Figure 5.8: Representative Motivation.

- Representative Reason: it is possible that the classifier space H considered for the problem does not contain the optimal classifier f. However, the combination of several classifiers can extend the space of H classifiers. In this way, the true classifier f can be approximated outside the space of classifiers. Figure 5.8 gives an illustration of this situation, where the optimal classifier f is outside the space of classifiers H.

These three reasons show the advantage of combining several classifiers and the limitations of individual classifiers. There are three possible combination approaches: the sequential approach, the parallel approach and the hybrid approach.

- Sequential approach: in this architecture the classifiers are trained sequentially, so that each classifier uses the results derived from the previous classifier (Figure 5.9). Therefore, at each stage, there is only one active classifier. There are two approaches to this sequential combination, the class set reduction approach and the reassessment approach. In the first approach, the number of possible classes is permanently reduced, while the second approach requires re-evaluation of the models, which are rejected in the previous step. The decision of a classifier in a serial combination is rejected if its confidence level is below a predefined threshold. Such an approach can be seen as a progressive filtering of decisions. Usually this will reduce the overall chain error rate. However, a combination of this type is particularly sensitive to the order in which the classifiers are placed. Indeed, even if they do not need to be the most efficient, the first classifiers in the chain must be robust.

- Parallel Approach: in this architecture the set of classifiers are trained in parallel independently of each other, and their results are then combined to give the final decision (Figure 5.10). In studies of the search

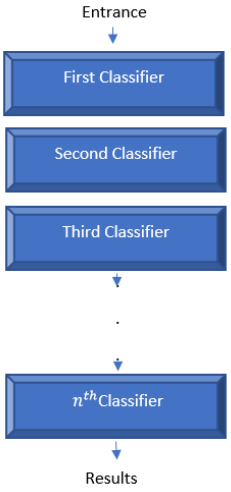


Figure 5.9: Sequential combination.

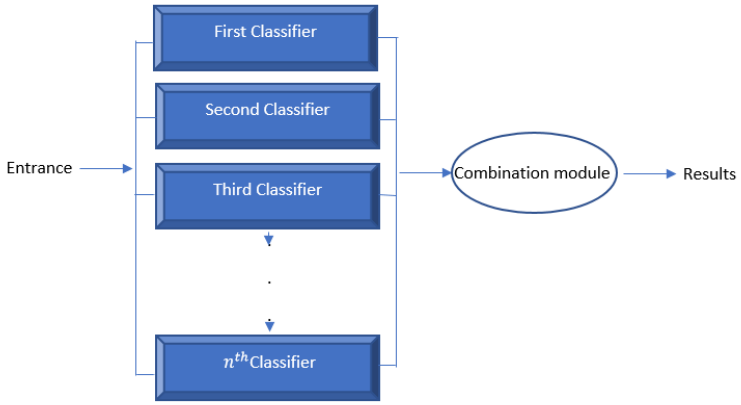


Figure 5.10: Parallel combination.

for combining classifiers, the parallel approach has gained a lot of attention because it has the advantage of being simple and easy to use. Unlike the sequential approach, the parallel organization classifiers requires that individual classifiers simultaneously produce their outputs. All of these outputs are then merged with a combination operator, such as a simple majority vote, to produce a final decision. In this approach the order in which the classifiers are placed has no influence.

- Hybrid Approach: The idea of the hybrid approach is to combine the two approaches above in order to retain the advantages of both. Many research works encourage the adaptation of the combination of classifiers to improve the performance of a model, or reduce the probability of selecting a weak classifier [253, 65].However, among these three approaches, the one that has aroused great interest in the scientific community is the parallel combination. The main motivation for this combination is to exploit the independence between classifiers which leads to the reduction of error in averaging. In the parallel category two approaches are proposed: the methods which use heterogeneous classifiers, that is to say classifiers of different types, leading to heterogeneous sets. We can cite the work of [253] who combined three heterogeneous classifiers, neural networks, the Bayesian classifier and hidden Markov models for a handwriting recognition application. There are also methods which use homogeneous classifiers, that is to say classifiers of the same types, leading to homogeneous sets. This therefore consists in applying the same

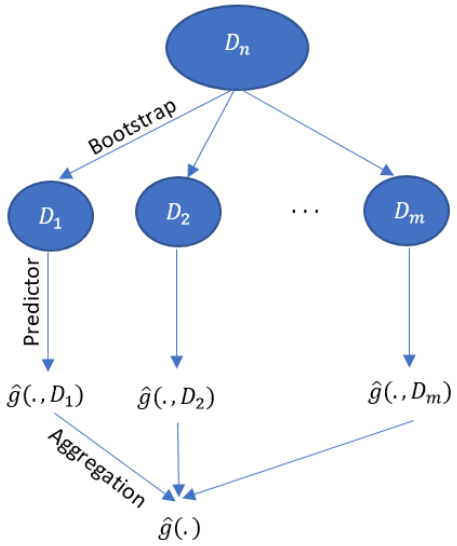


Figure 5.11: Bagging principle.

learning method for all the classifier but producing differences in their outputs. This amounts for example to selecting different training data subsets or different characteristic sub spaces for these data. We can cite an example that belongs to this family, it is that of the decision-making tree forests method [35]. This method uses a set of homogeneous classifiers which are the decision trees and applies the Bagging algorithm in order to create differences at the level of the predictions of each tree, which makes this model more generalized and therefore more efficient.

5.3.2 Decision tree forests

The decision tree forest method consists of a set of decision trees built in parallel, where each tree grows randomly and independently of the others. Randomness is important when building a tree. This ensures variety in the forest or, in other words, the trees become less correlated with each other. The RDF algorithm relies on the use of two principles of randomization: (a) Bagging and (b) selection of features to divide each node of a tree. The principle of the Bagging method: This is a method introduced by Breiman (1996). The word bagging is the combination of the words Bootstrap and Aggregating. The Bagging principle consists in building several bootstrap subsets from the training set D_n (Figure 5.11). Each bootstrap sample D_i, $i = 1,\ldots,m$ is obtained by a random selection and with replacement of n observations in D_n. From each boostrap D_i a classifier $\hat{g}(.,D_i)$ is induced. Finally, the collection of predictors is then aggregated by simply doing a majority vote. The resulting classifier model reduces the variance of individual classifiers [26].

Random selection of characteristics: For the selection of a subset of variables, the idea is to draw at each node k characteristics of the set of p characteristics available in a random manner and without replacement, $k << p$. At each node we select the best cut on the basis of k chosen variables, knowing that k is the same for all the nodes of all the trees of the forest. To choose the separation variable in a node, the algorithms test the different possible input variables and select the one that optimizes the node with respect to a measure of purity such as the information gain [196] or the Gini index [36]. The random selection of a reduced number of variables at each stage of the construction of a tree significantly increases variability by necessarily highlighting other variables.

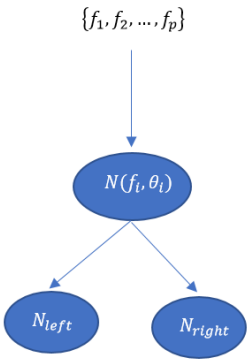


Figure 5.12: Principle of the selection of characteristics.

Each basic model is obviously less efficient but the aggregation ultimately leads to a more robust and precise classifier. This also makes it possible to further reduce the calculation time. Figure 5.12 illustrates the principle of random selection of features. So we can summarize the steps of the decision tree forest algorithm as follows:

1. For $i = 1, \ldots, m$

   (a) Randomly pull a bootstrap from D_n.

   (b) Build a tree t_i for each bootstrap D_i.

   - Randomly draw k characteristics among the p, with $k << p$.
   - Among the k characteristics, choose the one which gives the best division of the node.
   - Divide the node into two child nodes

   (c) Calculate the classifier on this sample: $\hat{g}(., D_i)$.

2. Output $\hat{g}(x) = argmax_c \Sigma_{i=1}^{m} I_{\hat{g}(., D_i)=c}$

OOB Error: In RDF method, there is no need for cross validation or a separate test set to get an estimate of the prediction error. It is internally estimated as follows: each tree is constructed using a different random bootstrap sample of original data. Each bootstrap leaves out a certain number of observations ($\approx 1/3$), called OOB (Out Of Bag) observations which are used to estimate the performance of the model by calculating the Out-of-bag error (errOOB). Therefore, for a given tree T_b in the forest, only 2/3 of the training data is used in its construction and the remaining 1/3 presents the associated OOB samples. For each observation (x_i, y_i), we select the bootstrap samples D_i not containing (x_i, y_i). For these samples, we say that the observation (x_i, y_i) is an OOB sample. We predict this observation with all trees built on these bootstrap samples. We aggregate their predictions by a majority vote and we denote by $\hat{y}_i$ the OOB prediction of x_i. Subsequently, we can calculate the average OOB error rate of the forest with the following formula:

$$err_{OOB} = \frac{1}{n} \Sigma_{i=1}^{n} I_{y_i \neq \hat{y}_i} \tag{5.11}$$

The advantage of the OOB error is that the original and entire set of data can be used to build the RDF classifier. Unlike cross-validation methods in which a subset of the samples is used for the construction of an

RDF, the OOB procedure allows all the samples to be used for the construction of the classifier. This results in RDF classifiers that have greater precision than that obtained from cross-validation. Another benefit of using the OOB error is the computation time, especially when it comes to a number large data volumes, where building a single RDF can take days or even weeks. With the OOB procedure, only one RDF must be built, unlike the cross validation of k groups where k RDFs must be built [42, 298]. The OOBs samples will therefore be used for the internal evaluation of a forest and also for the estimation of the importance of the variables. Importance of variables: Following the RDF principle, the importance of a variable f is the difference between the prediction precision (i.e. the number of correctly classified cases) before and after the permutation of that variable in the OOBs samples, averaged over all trees. A large decrease in the accuracy of the prediction denotes the importance of this feature. The calculation of the prediction error of the OOBs samples makes it possible to estimate the importance variables in the following way: the error that each tree commits on its associated OOB sample is calculated. In all OOBs samples, the values of the variable f are randomly permuted. The error that each tree commits on its permuted OOB sample is recalculated and subsequently compared with the original OOB error (before permutation). If the error rate of OOB after the permutation is larger than that with the original OOB observations, the variable f is considered important. We summarize the steps of calculating the importance of a variable f by RDF approach as follows:

1. For each tree $t = 1 \dots T$ in the forest, calculate err_{OOBt}, the average error rate over all OOB observations in tree t.

$$err_{OOBt} = \frac{1}{Card(OOBt)} \Sigma_{x_i \in OOBt} I_{y_i \neq \hat{y}_{it}} \tag{5.12}$$

   OOB_t contains the observations that do not appear in the bootstrap sample used to build the tree t, Card $(OOBt)$ denotes its cardinality. y_i and $\hat{y}_{it}$ present respectively, the true label and the prediction of the i^{th} observation by the tree t.

2. Randomly permute the feature velaues, f, in the OOB_t sample. This gives a disturbed sample, denoted $OOBt^f$. Calculate in n $errOOB_t^f$, the average error rate on $OOBt^f$.

$$err_{OOB_t^f} = \frac{1}{Card(OOB_t^f)} \Sigma_{x_i \in OOB_t^f} I_{y_i \neq \hat{y}_{it}} \tag{5.13}$$

3. The importance of a features f by a tree t is calculated as follows:

$$I^t(f) = err_{OOB_t^f} - err_{OOB_t} \tag{5.14}$$

   Note that $I^t(f) = 0$, if the features f is not in the tree t. The score of a features f is then calculated by the average of the importance on all the trees.

$$I(f) = \frac{1}{T} \Sigma_{t=1}^{T} I^t(f) \tag{5.15}$$

   where T is the number of trees.

Thus, the more the random permutations of the features f lead to a large increase in the error, the more the

characteristic is considered important.

5.3.3 Experimental results of the classification of expressive gestures with the RDF method

In the first experimental part of this study, we evaluate the performance of our global descriptor on the characterization of the different expressive gestures of our XEM data set. So, we consider the entire data set, 1100 sequences (11 participants 5 gestures 4 emotions 5 repetitions). The first step is to adjust the parameters of the RDF method. We adjust the two most important parameters in the RDF method, which are: the number of trees T and the number of features selected for each divisionc_{max}. We use two different validation methods: 3-fold cross-validation and estimation of the average OOB error rate. We measured the accuracy of the models with F-score as follows.

$$F-score = \frac{2}{\frac{1}{recall} \times \frac{1}{precision}} = 2 \times \frac{precision \times recall}{precision + recall} = \frac{tp}{tp + \frac{1}{2}(fp + fn)} \tag{5.16}$$

In which

- Precision is the fraction of true positive examples among the examples that the model classified as positive.
- Recall, also known as sensitivity, is the fraction of examples classified as positive, among the total number of positive examples.
- tp is the number of true positives classified by the model.
- fn is the number of true positives classified by the model.
- fp is the number of false positives classified by the model.

And

- tp refers to the number of instances which are relevant and which the model correctly identified as relevant.
- fp refers to the false positive rate, that is the number of instances which are not relevant but which the model incorrectly identified as relevant.
- fn refers to the number of instances which are relevant and which the model incorrectly identified as not relevant.

Therfore, as the results we have:

- With the cross-validation method: for the parameter T, we vary its value from 10 up to 300 trees. For the c_{max} parameter, we test 3 values: 85, $\sqrt{85}$ and $\log_2(85)$. The best average F-score is $0:83$, obtained for $c_{max} = \log_2(85)$ and $T = 150$.
- With the OOB error measurement: we measure the OOB error rate (err_{OOB}) for each possible value of the torque (T, c_{max}). Figure 5.13 shows the results obtained from the OOB error rate for each variation of T and c_{max}. The green, blue and orange curves correspond respectively to the values of c_{max} equal to

85, $\sqrt{85}$ and $\log_2(85)$. We notice that the orange curve ($c_{max} = \log_2(85)$) is the lowest, it is the curve which displays the lowest values of err_{OOB}. We also notice that the two blue ($c_{max} = \sqrt{85}$) and orange ($c_{max} = \log_2 85$) curves are very close with a slight superiority to the orange curve. We have identified the value of T where the OOB error rate stabilizes with a minimum value (around 0.06). We found a value of about 150 trees, which confirms the result obtained with the 3-group cross-validation method.

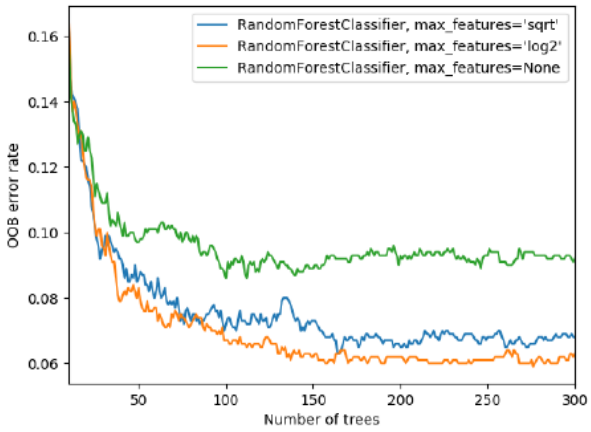


Figure 5.13: Variation of the OOB error rate (err_{OOB}) as a function of (T, c_{max}).

We have also presented the confusion matrix of all expressive gestures in Figure 5.14.

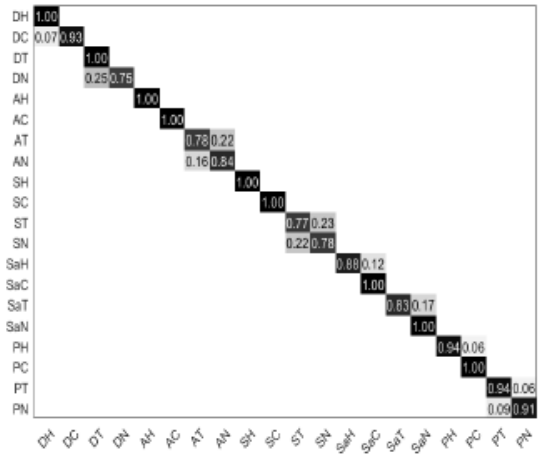


Figure 5.14: Confusion matrix of expressive gestures, 5 gestures (D dance, A Move forward, S Make a sign, Sa stop and P point) performed with 4 states (H happy, C angry, T sad and N neutral).

In this case, we have 20 classes (5 gestures 4 emotions). As we can see, the highest values are marked in the diagonal of the matrix, which confirms the performance of our results. Following the matrices, there is some confusion in the same gesture when it is performed with different emotions. For example, in the gestures "move forward", "make a sign" and "point" we find a confusion between the state "sad" and the state "neutral". Generally, we can say that our motion descriptor manages to characterize the quantitative as well as the qualitative aspects of the movement. The second test, we want to classify the emotions according to the type of gesture. We applied the RDF method with its adjusted values. The database is divided into 5 groups per class gesture. Each group is made up of 220 sequences (1 gesture 4 emotions 11 people 5 repetitions). The same 3-group cross-validation method is applied. Table 5.2 shows the results obtained in the recognition of emotions

Gestures	Waving	Moving Forward	Dancing	Stopping	Pointing	All gestures
F-Scores	0.93	0.80	0.75	0.84	0.91	0.83

Table 5.2: Results of the recognition of emotions in the different gestures of our data set

in each group of gestures considered (dancing, moving forward, waving, pointing and stopping). The matrices of the confusions of different emotions are presented in Figure 5.15. The results show that our descriptor succeeded in distinguishing between the different emotions in each type of gesture. Confusion matrices between expressed emotions.

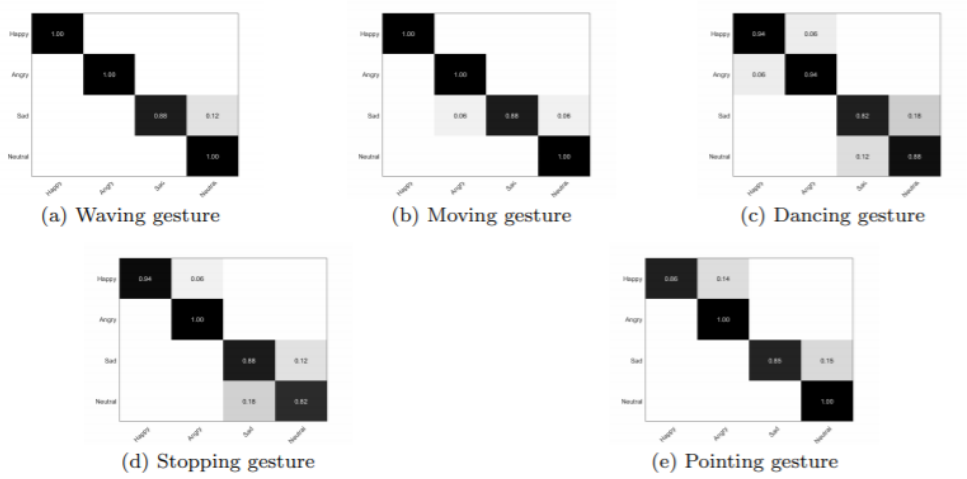


Figure 5.15: Confusion matrices between expressed emotions (in rows) and perceived emotions (in columns) for each gesture using RDF method.

Some emotions have been recognized successfully, for example the emotions of joy and anger in "waving" and "moving" actions have been 100% recognized. However, some confusions were obtained in the same movement, in particular between the emotions of sadness and neutral, as in the gestures "to dance" and "to stop". Based on these results, we can endorse the effectiveness of our motion descriptor in recognizing human movements and emotions.

5.3.4 Selection of relevant characteristics with the RDF method

Feature selection methods

Feature selection is an important topic in several areas including, in pattern recognition, exploratory data analysis, especially for large data sets. Feature selection (also known as subset selection) is a process commonly used in machine learning, which involves selecting a subset of the most discriminating features from among the set of available features. The best subset contains the most relevant characteristics that contribute the most to have the best accuracy. This helps identify and remove irrelevant and redundant information as much as possible. The feature selection algorithms can be divided into three approaches: Filter, Wrapper and Embedded.

- The Filter model selects characteristics based on a performance measure. The selection process is independent of the classification process. This model applies a statistical measure to assign a score to each characteristic and provide a ranking. From the ranked list, the characteristics with the highest scores are selected and the characteristics with the low score are removed. These characteristics can be chosen

manually or by defining a threshold. This type of approach is therefore based solely on the properties of the data. It is independent of any particular machine learning algorithm. Examples of this method are the information gain [99], the gain ratio [278], the Chi-square test [278], the coefficients of correlation [291], etc. These methods are fast and independent of the classifier. On the other hand, their major drawback is that they ignore the impact of the selected subsets on the performance of the learning algorithm.

- The wrapper model requires a predetermined training algorithm to use its performance as a criterion for evaluating the set of selected characteristics. This type of approach generates subsets of characteristics and evaluates them using a classification algorithm. This evaluation is repeated for each subset, and a classification algorithm call is made for each evaluation. The generation of each subset depends on the chosen search strategy. Two methods are proposed in the wrapper approach, the backward selection method, a search algorithm introduced by [158] and the forward selection algorithm by [275]. In the case of the backward search, the method starts with the set of all characteristics and removes the characteristics one after the other. At each step this method removes the feature that has the highest error until any further removal greatly increases the error. In this "top-down" search, the ignored characteristics cannot be selected again. In the case of forward search, the method starts without variables and adds the characteristics one after the other. At each step, it adds the characteristic which has the minimum error until any further addition means no decrease in error. In this "bottom-up" search, the selected characteristics cannot be rejected later. Wrapper methods essentially solve the "real" problem (optimizing classifier performance), but they are also more computationally expensive compared to filtering methods because of the repeated training and cross-validation steps.

- The Embedded model: in this approach, the selection of characteristics is an integral part of the classification model. These methods assess the characteristics that best contribute to model accuracy during model creation. We can cite the example of the RDF method explained previously which makes it possible to measure the importance of variables during the creation of trees.

5.3.5 Our RDF-based feature selection method

In our work, the main objective is to select the optimal subset of characteristics for the characterization of the emotional expression of the body in the control gestures of our data set. As we have seen, the RDF method has shown its performance in the learning stage and also its ability to measure the relevance of the characteristics based on the measurements of the error rates of the OOBs samples. So our idea is to exploit the advantages of this method and consider it as an embedded model for feature selection. It makes it possible to measure the importance of characteristics during the creation of trees. So, our important feature selection algorithm is considered as follows: we train the RDF model with all features and we record the corresponding OOB error rate. Then, we measure the importance of each characteristic and we sort them in decreasing order of importance. We recursively remove the less important features. Here in this step the less important characteristics and also those which are redundant are removed by applying the Tukey HSD (Honestly Significant Difference) test. So, following each sort, we remove the set that does not contribute to a significant change in the OOB error following Tukey's test ($\alpha = 0:05$). This makes our system more efficient and faster. The process stops when the number

of characteristics remaining in the set is equal to 1. The output of this algorithm gives a curve which describes the decrease in the system error (OOB error rate) as a function of the subsets of selected characteristics. The subset that corresponds to the minimum value of the OOB error is considered the optimal subset. Algorithm 3 summarizes the different steps of our proposed feature selection algorithm:

Algorithm 3 Feature selection process.

1. Input: $v_0 = \{f_i\}, i = 1,\ldots,p$ ► v_0 is the whole feature set.
2. Output: $v^* = \{f_i\}, i = 1,\ldots,p^*$ ► v^* subset of most relevant features.
3. $k = 0$.
4. While $|v_k| \geq 1$ do
5. Compute and record OOB error rate: $E_k(v_k)$
6. for $i = 1$ to p do
7. Compute $I(f_i)$ ► Importance of each feature in v_k.
8. end
9. Sort $\{f_i\}$ in descending order according to values of $I(f_i)$
10. $f_{min} = argmin_i\{I(f_i)\}$
11. Apply Turkey's test and select set of features $\{f_t\}$ that does not lead to a significant changement of E_k.
12. $R = f_{min} \sum\{f_t\}$.
13. $v_{k+1} = \frac{v_k}{R}$
14. $k = k + 1$
15. end
16. $v^* = argmin_l\{E_l(v_l)\}$.

5.3.6 Results of relevant characteristics with RDF

In this part we carry out two experiments, first we evaluate the relevance of the characteristics of our movement descriptor in the characterization of the whole set of expressive gestures of our data set. In the second experiment, we study the importance of characteristics to each emotion. For this we apply our selection algorithm in each gesture of the data set. In order to define the most relevant characteristics for each emotion, we take the neutral state as a reference each time. The first study consists in considering the whole data set, so the RDF model is trained on all the learning gestures of the data set. We apply feature selection algorithm 3 on our data set. The results are shown in Figure 5.16 which displays the OOB error curve as a function of the selected characteristics. The minimum value of the error rate of OOB is 0.04 obtained with the whole set of characteristics (85 characteristics). This confirms the importance of combining all components of LMA for the characterization of our data set of expressive gestures.

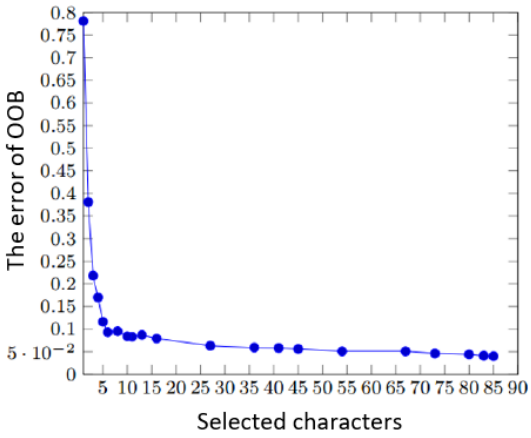


Figure 5.16: Variation of the OOB error rate according to the characteristics selected.

In the second experiment, we divide the database into 5 groups by class of gesture. So, each time we consider the same gesture made with several emotions. We assess the importance of each characteristic for the characterization of each emotion by comparing with the neutral state. Therefore, the RDF model is trained each time on two classes of gestures, a gesture made with an emotion (joy, anger, or sadness) and the same gesture with the neutral state. In this study we were more interested in the qualitative aspect of the movement, since it is the same gesture performed with several emotions, so we will necessarily have the same importance between the quantitative characteristics. For this, we keep only the factors of the two components Form and Effort which describe the different expressive qualities perceived in the movements. The relevance of each characteristic is calculated with the same selection algorithm used before. The objective of this study is to find the most relevant characteristics to characterize each emotion according to the type of gesture. Table 5.3 summarizes the results found for each gesture and each emotion.

The application of our feature selection algorithm shows that the factors of the Effort component, particularly the qualities of time and weight, involve a strong discrimination between the neutral state and the emotions of joy and anger. So, we can say that the emotions of joy and anger are characterized by the speed and force of the movement. While the emotion of sadness is characterized by the two components of LMA (Effort and Shape), with the exception of the factor of directional movement.

5.4 Characterization of expressive gestures with the human approach

5.4.1 Evaluation of emotions

The evaluation of emotions with human perception can be done in two forms:

- Self-Evaluation: A self-evaluation is a test, measure, or survey that is based on the participant's own report of their feelings, attitudes, or beliefs. Self-assessments are commonly used in psychological studies, largely because a lot of valuable and diagnostic information about a person is revealed to a researcher or clinician based on a person's report on themselves. One of the most commonly used self-assessment tools is the Minnesota Multiphasic Personality Inventory (MMPI) for personality testing. This is a personality

Expressive motions		Motion descriptor					
		Shape		Effort			
		Shape flow	Shaping	Time	Weight	Space	Flow
Happy	Moving			std_vl range_vl std_vr range_vr	mean_al std_al, mean_ar std_ar range_ar std_ah range_ah		
	Dancing				mean_al std_al std_ar mean_ah std_ah		
	Waving			std_vl range_vr	std_al std_ar		
	Stopping			mean_vl std_vl range_vl mean_vr std_vr range_vr std_vh range_vh	mean_al std_al range_al mean_ar std_ar range_ar mean_ah std_ah range_ah mean_as std_as range_as		
	Pointing			range_vl std_vr range_vr std_vs range_vs	mean_al mean_ar std_ar range_ar mean_ah std_ah range_ah mean_as std_as range_as		
Angry	Moving			range_vl,	std_al std_ar,range_ar		
	Dancing				std_al range_al std_ah		
	Waving			range_vr	mean_al std_al range_al mean_ar std_ar range_ar		
	Stopping				mean_ah std_ah mean_as		
	Pointing				std_ar,range_ar		
Sad	Moving	mean_V std_V range_V	D_H D_F	mean_vl std_vl mean_vr,std_vr,	mean_al mean_ar mean_ah mean_as	Sl Sr	range_pitchr range_yawr range_pitchh range_yawh
	Dancing	std_V		std_vr,range_vr	mean_al std_al range_al std_ar mean_as std_as	Sr	range_pitchl range_yawr range_pitchh
	Waving		D_H D_S	mean_vl std_vl range_vl mean_vr std_vr range_vr	mean_al std_al mean_ar std_ar		
	Stopping	std_V range_V,	D_F	mean_vl std_vl range_vl mean_vr std_vr range_vr mean_vh std_vh std_vs range_vs	mean_al std_al mean_ar std_ar mean_as std_as	Ss	range_pitchr
	Pointing	mean_V range_V		mean_vr, std_vr,range_vr	mean_ar std_ar range_ar	Sr	range_pitchh range_yawh

Table 5.3: The relevant characteristics for each emotion through the different gestures.

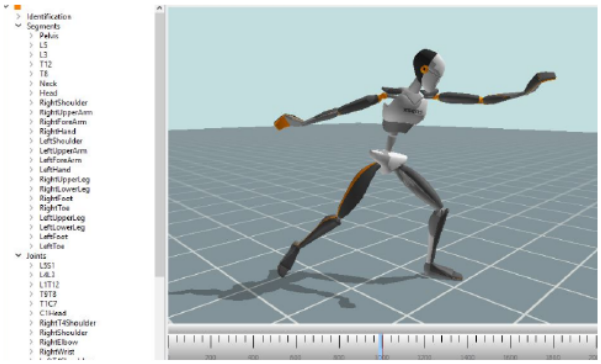


Figure 5.17: Reproduction of gestures with an avatar.

self-questionnaire for diagnostic, descriptive and therapeutic purposes. This tool has the advantage of being less expensive in time than the observation methods. It can reach many more test subjects so that a researcher can get results in a matter of days without having to observe a population during periods that may last longer. However, collecting information through a self-assessment report has its limitations. Indeed, the results are often biased when people report their own experiences by several factors. For example, they may give the most socially acceptable answer rather than being truthful. They may also be unable to assess themselves properly. The wording of questions can be confusing or have different meanings for different topics. Sometimes in the field of medicine, these studies can in some cases have problems of validity. Patients may exaggerate symptoms to make their condition worse, or underestimate the severity or frequency of symptoms to minimize their problems.

- Observer evaluation: this is an evaluation carried out by a set of observers on the behaviors, symptoms, or attitudes of other people. The observer response format has often been based on the forced choice option, assigning a single tag to expressed bodily behavior is required and the most frequent tag is used. Observer-based evaluations have shown advantages over self-evaluations especially in the psychological domain in the evaluation of personality [57, 193]. There are several reasons for this advantage, we can cite the reliability in the assessments (i.e higher Cronbach alpha [20]). Moreover, this reliability is a necessary condition for validity [176].

In our case, we have chosen the second type of evaluation, given the limitations of the first method. We are more interested in the reliability of the test than in its duration.

Participants

We call on 10 observers (5 men and 5 women) from the University of Evry Val d'Essonne whose ages vary between 28 and 37 years (average = 30.9 years, standard deviation = 3.16). Each participant is invited to watch the recorded videos and rate the emotion expressed in each gesture using the 5-item Likert scale (from 1 = strongly disagree, 3 = neutral, to 5 = strongly agree). To perform a reliable evaluation, all expressive gestures recorded in the videos are reproduced by a virtual avatar (see Figure 5.17). It helps observers to evaluate emotions without being influenced by certain factors like facial expressions, gender, sex, etc.

Inter-observer reliability in the evaluation of emotions

After the evaluation by the various observers, a very important step must be carried out for the validation of

Figure 5.18: Inter-Viewer reliability of the emotions perception using the Cronbach's alpha.

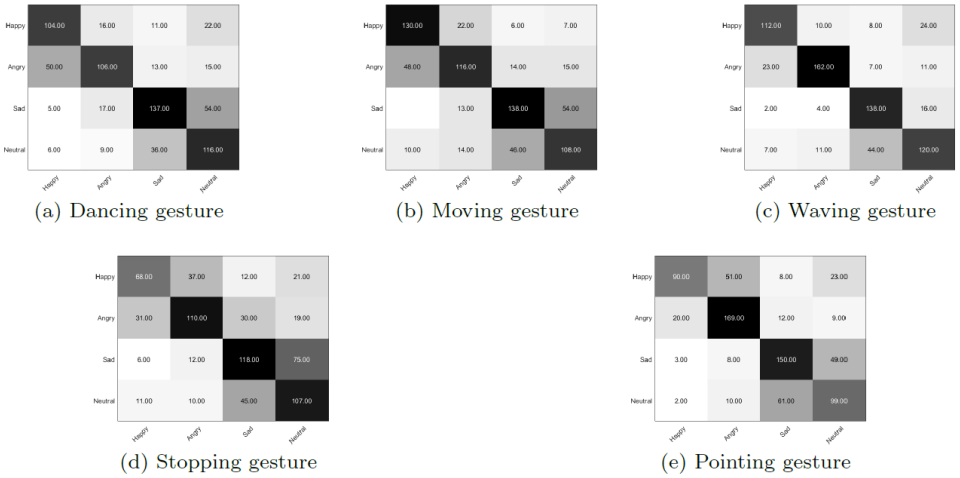
the experimental part, which is the estimation of the degree of homogeneity and of cohesion among observers. This study makes it possible to identify the parts which contribute little to the evaluation. The reliability index is then expressed by measuring the consistency between the different observers. The measurement is said to be reliable if there is a sufficiently high degree of agreement between the different evaluations and is not reliable otherwise. Several statistical indicators make it possible to evaluate the inter-element agreement, the most used is the Cronbach coefficient, given by the following formula:

$$\alpha = \frac{n}{n-1}(1 - \frac{\Sigma_{i=1}^{n} V_i}{V_T}) \tag{5.17}$$

where n is the number of items, V_i refers to the variance of scores on each item and V_T is the total variance of all scores. The Cronbach alpha is interpreted as a classic correlation coefficient, the closer it is to 1, the more reliable the score. We calculate this coefficient to measure the correlation between the different scores given by the observers for each emotion. As shown in Figure 5.18, the Cronbach coefficient is always greater than 0.8: for happy (0.897), angry (0.894), sad (0.879) and neutral (0.814). This means, according to [245], that there is great consistency between the different evaluations made by observers in the perception of emotions.

Results

For the classification of emotions with the human approach, we take the resulting ratings from observers, and we consider a recognized emotion if the given score is greater than 3 (neutral state). In order to classify the emotions, we take a gesture each time and for each emotion we calculate the number of times the score is greater than 3. The results are presented in the confusion matrices in Figure 5.19. The rows correspond to the emotions expressed and the columns to the emotions perceived (evaluated) by the observers. Diagonal cells correspond to the number of times the emotions reported by observers (score$>$ 3) are recognized. Non-diagonal cells contain the frequency of misclassifications. As we can notice, the highest values are in the diagonal in all gestures, with greater confusions between the emotions of joy and anger and between the emotion of sadness and the neutral state. In a second experiment, we consider all the expressive gestures and we calculate the average of the scores of the observers in each expressed emotion. As we can see in Figure 5.21, the emotion of



(a) Dancing gesture (b) Moving gesture (c) Waving gesture

(d) Stopping gesture (e) Pointing gesture

Figure 5.19: Confusion matrices between expressed emotions (in rows) and perceived emotions (in columns) for each gesture based on viewers rating.

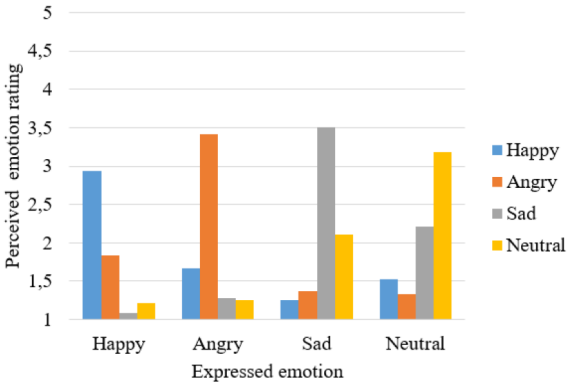


Figure 5.20: Mean ratings of emotions perception by 10 viewers.

joy has been successfully recognized by observers with some confusion arisen with the emotion of anger. The emotion of joy is weakly confused with that of sadness. For the emotion of anger, the highest mean score was obtained in the perception of the emotion of anger followed by the emotion of joy. The neutral and sadness states were well recognized by observers in the various gestures with a bidirectional confusion between the two emotions by some observers.

5.4.2 Selection of characteristics with the human approach

For the second experiment concerning the evaluation of our descriptor with the human approach, we ask another group of 10 observers from the University of Evry Val d'Essonne, their ages vary between 28 and 31 years old, to look at the even videos and assess the components of LMA. Likewise, in this part, we have focused on the factors of the two Effort-Shape components (form flow, shaping, directional movement, space, time, weight and flow), because they are the only ones responsible for the specification of expressive human movements. We use the 7-item Likert scale as follows:

- Shape flow: volume of the convex envelope of the skeleton (from 1 = very small to 7 = very large).
- Shaping: distance between the center of the skeleton and the extremities (from 1 = very contracted to 7 = very extended).

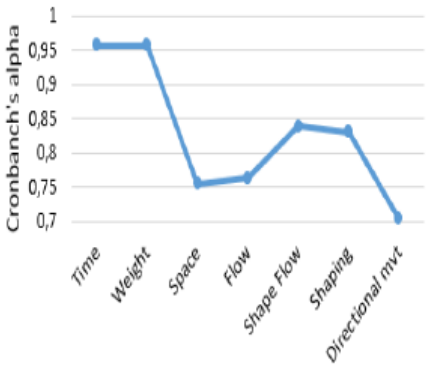


Figure 5.21: Inter-Viewer reliability of Effort-Shape features rating using Cronbach's alpha.

- Directional movement: the way of the joints of the body (from 1 = very curvilinear to 7 = very straight)
- Time: the speed of the movement (from 1 = very sustained to 7 = very sudden).
- Weight: the force of the movement (from 1 = very light to 7 = very strong)
- Space: straightness of movement (from 1 = very indirect to 7 = very direct)
- Flow: flexibility of movement (from 1 = very free to 7 = very close)

For example, to evaluate the shape flow factor it is necessary to perceive the development of the volume of the convex envelope of the skeleton. If the observer perceives that the volume of the skeleton has increased sharply throughout the movement, he gives the form flow factor a score of 7 and if he perceives a sharp decrease in volume, he gives the score of 1.

Inter-observer reliability in the evaluation of characteristics

Likewise, this evaluation requires the measurement of the reliability between the scores given by the observers in the evaluation of the characteristics. According to the results presented in Figure 5.21, the Cronbach coefficient is greater than 0.7 (accepted level) in all evaluations. This confirms consistency between observers when evaluating characteristics in our expressive base. The highest reliability coefficient is 0.958 obtained in the evaluation of the factors of time and weight. This result shows a strong homogeneity between the observers in the estimation of these two factors. However, there is a less important but acceptable correlation in the evaluation of the factors of space, flow and directional movement, respectively. with Cronbach coefficients of $0.754, 0.763$ and 0.703. For the shaping factors and the shape flow, there is good agreement between the scores given by the observers, respectively with Cronbach coefficients of 0.830 and 0.839.

To study the importance of characteristics in the characterization of human emotions, we measure the correlation between the assessments made in the perception of emotions and in the characterization of LMA factors. This helps us to define the relationship between descriptor characteristics and the 4 states (happy, angry, sad and neutral). We collect the scores given to the factors of Effort-Fitness and the perception of emotions and we calculate the Pearson correlation coefficient. Pearson coefficients are used in statistics to measure the relationship between two variables, called r Pearson, with a value between +1 for a perfectly negative correlation and +1 for a perfectly positive correlation, 0 if the two variables do not represent any correlation. Pearson's

correlation between the variables X and Y is calculated by the following formula:

$$r_{XY} = \frac{\Sigma_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\Sigma_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\Sigma_{i=1}^{n}(Y_i - \bar{Y})^2}} \tag{5.18}$$

Table 5.4 summarizes the results of the Pearson coefficients obtained in this study: For the emotion of joy, we find a positive correlation with the factors of Form (Shape flow $r = 0.541$, $p < 0.001$, Shaping $r = 0.542$, $p < 0.001$, Directional Movement $r = 0.269$, $p < 0.001$) and the Effort factors (Time $r = 0.555$, $p < 0.001$, Weight $r = 0.543$, $p < 0.001$). So the emotion of joy is significantly characterized by an increase in Shape and an extension of the limbs of the body. This emotion is also associated with the speed and strength of movement. In [159], also the authors confirmed in their experience the same relationship between the emotion of joy and the qualities of Effort.

		Happy	Angry	Sad	Neutral
Shape	**Shape Flow**	**0.541****	0.132	**-0.524****	-0.316**
	Shaping	**0.542****	0.194**	**-0.613****	-0.328**
	Directional mvt	0.269**	**0.568****	**-0.505****	-0.397**
Effort	**Time**	**0.555****	**0.622****	**-0.795****	**-0.554****
	Weight	**0.543****	**0.640****	**-0.780****	**-0.594****
	Space	0.326**	**0.682****	**-0.566****	-0.487**
	Flow	-0.316**	**-0.665****	**0.559****	0.497**

Table 5.4: Pearson's correlation coefficients r between Effort-Shape factors and expressed emotions ratings (**. correlation is significant at the 0.001 level).

For the emotion of anger, we find a strong positive correlation with the following three Stress factors (Time $r = 0.622, p < 0.001$, Weight $r = 0.640, p < 0.001$, Space $r = 0.682, p < 0.001$) and a positive correlation with the factor of directional movement ($r = 0.568, p < 0.001$). The flow factor is negatively associated with the emotion of anger ($r = -0.665, p < 0.001$). So the emotion of anger is strongly characterized by fast, strong, direct, straight, free movement. Some qualities are consistent with those found by [159]. The authors found a positive correlation between the emotion of anger and the qualities of speed and strength of movement. For the emotions of joy and anger, we found a similar relationship with the factors of the Effort-Fitness components, but with a different correlation importance. In the assessment of the qualities of Effort, the emotion of anger was rated as significantly faster, stronger and freer than the emotion of joy. However, in the evaluation of Form factors, the emotion of joy was evaluated by a more developed form. The emotion of sadness is negatively correlated with all of the Stress-Form factors, except the flow factor: form factors (form flow $r = -0.524, p < 0.001$ shape $r = -0.613, p < 0.001$; directional movement $r = -0.505, p < 0.001$), effort factors (time $r = -0.795, p < 0.001$; weight $r = -0.780, p < 0.001$, space $r = -0.566, p < 0.001$, Flow $r = 0.559, p < 0.001$). According to the shape factor assessment, the emotion of sadness was significantly characterized by a narrowed shape, contracted extremities of the body, and bent movement. According to the Effort Factor assessment, the emotion of sadness is characterized by light, bound, sustained, and indirect movement. The neutral state, had a relationship with the qualities of Effort-Shape similar to that obtained with the emotion of sadness but with a less strong correlation. Likewise, [159] found that both emotions (relaxed and sad) correlate with the same qualities of

Effort (slowness and weakness).

5.5 Expressive global descriptor inspired by geometric and time dependent features

In this section, we tried to improve the results of our work by changing the constituent elements of the descriptor as well as the method of classifying and identifying emotions. To do this, we used some qualitative and quantitative characters that are derived from the 3D coordinates of joints of body to describe a gesture. The features extracted in this section are divided into two categories: time-dependent and geometric characters, which we will explain below.

5.5.1 Geometric Characters

- $EA = (ea_1, \ldots, ea_n)$: Euler angle extraction of all the joints. $1 \leq i \leq n$ where n is the total number of joints

- $\Theta = (\theta_l^1, \theta_r^1, \theta_l^2, \theta_r^2, \theta_l^3, \theta_r^3, \theta_l^4, \theta_r^4)$: The angles around elbows (θ_l^1, θ_r^1), neck (θ_l^2, θ_r^2), hips (θ_l^3, θ_r^3) and knees (θ_l^4, θ_r^4) which are computed by:

$$\theta^{j_j} = \arccos \frac{\overrightarrow{j_j - j_i} \cdot \overrightarrow{j_k - j_j}}{\left\|\overrightarrow{j_j - j_i}\right\| \left\|\overrightarrow{j_k - j_j}\right\|} \tag{5.19}$$

  Where $j_i = (x_i, y_i, z_i)$, $j_j = (x_j, y_j, z_j)$, and $j_k = (x_k, y_k, z_k)$ $3D$ coordinates of three consecutive joints.

- $D = (d_h, d_{hf}^l, d_{hf}^r, d_{hn}^r, d_{hn}^l, d_f)$: The distances between the two hands (d_h), between the left hand and left foot (d_{hf}^l), between the right hand and right foot (d_{hf}^r), between the right hand and neck (d_{hn}^r), between the left hand and neck (d_{hn}^l) and between the feet, d_f. For example, the distance between the hands is calculated from the following formula:

$$d_h = \sqrt{(x_h^r - x_h^l)^2 + (y_h^r - y_h^l)^2 + (z_h^r - z_h^l)^2} \tag{5.20}$$

  Other distances are also obtained through Formula 5.20.

The geometric characters are shown in the figure 5.22. Another factor that can describe how a gesture is performed is the softness of an action. According to [86], "smoothness" is synonymous to "having small values of high-order derivatives". In this work, we use the velocity and acceleration to describe the trajectory of the movement. The hands are expected to have a higher curvature when they follow a circle, compared to when they follow a straight line [86]. The curvature measures how fast the unit tangent vector to the curve rotates. Curvature C calculate the rate at which a tangent vector turns as a trajectory bends. So that a path following a small circle will bend sharply, and then a higher curvature while a trajectory following a straight line will led to have a zero curvature. For each trajectory the curvature is calculated by:

$$C(j_h) = \frac{\|v(j_h) \times a(j_h)\|}{(\sqrt{(v(j_h)_x)^2 + (v(j_h)_y)^2 + (v(j_h)_z)^2})^3} \tag{5.21}$$

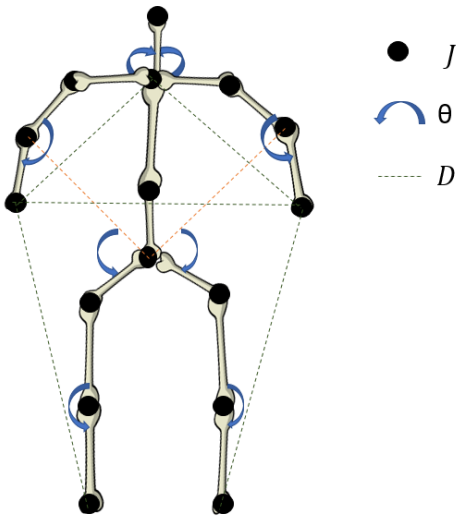


Figure 5.22: The geometric characters of a skeleton.

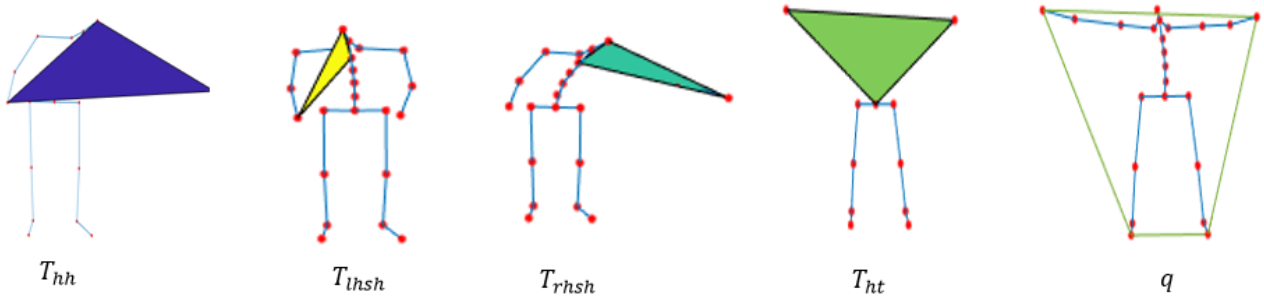


Figure 5.23: Geometric shapes formed by different parts of the body.

Where $v(j_h)$ and $a(j_h)$ represents the velocity and acceleration of right and left hand.The next component considered in this section to make a descriptor is the space occupied by the different parts of the body. This is a factor in explaining the expansion and contraction of body parts. To do this, we divided the body into several sections and calculated their area during the gesture. The divided sections are as follows:

- The triangle between two hands and the head, T_{hh}.
- The triangle between two hands and the torso, T_{ht}.
- The triangle between the right hand, the shoulder center and the head, T_{rhsh}.
- The triangle between the left hand, the shoulder center and the head, T_{lhsh}.
- The quadrilateral consists of two hands and feet, q.

These geometric shapes are shown in the figure 5.23. The last factor we have added to the descriptor is the minimum ellipse-shaped space occupied by the skeleton of the body in space. To do this, we used the algorithm presented in [168] to find the radius and center of the minimum 3D ellipse surrounded by the body skeleton and calculated its volume. In Figure 5.24, the ellipse obtained from this algorithm, SE, is shown in 2D and 3D view.

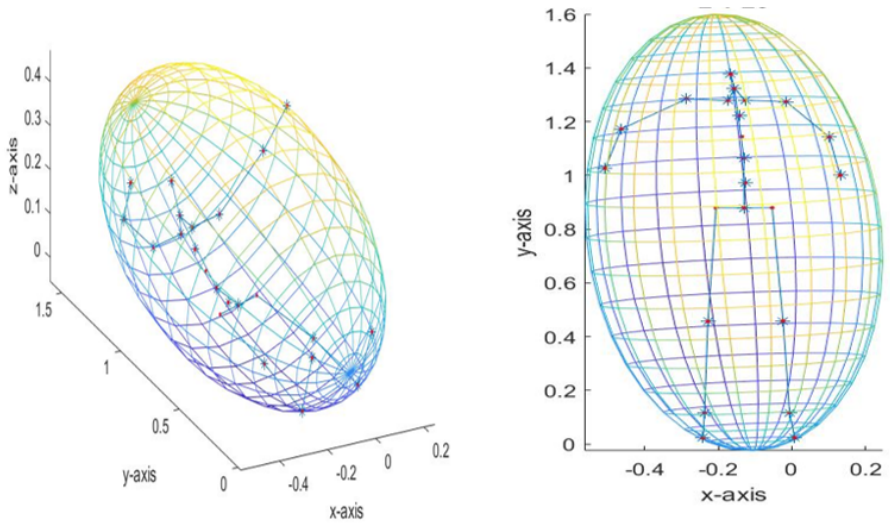


Figure 5.24: 3D and 2D view of surrounded ellipse, SE.

5.5.2 Time-dependent characters

To define a suitable and robust descriptor, we must select characters whose nature does not vary by changing variables such as light, location, and the person performing the gestures. Time variables are one of these variables that can weaken the descriptor and reduce the accuracy of identification due to changes in the descriptor. Because in general the speed of daily activities varies from person to person. In fact, these components show us how a gesture is made. A study by human observers to understand emotions while walking has shown that there are significant differences in walking patterns between people with different emotions [91]. In their pattern, they recognized happiness with a bouncy gait, sadness with slow gait, anger with a fast gait and fear with fast and short walks. Given the influence of time-dependent characters on recognizing emotions, we also used the following factors to construct our descriptor:

- $V = (v_1, \ldots, v_i, \ldots, v_n)$ $1 \leq i \leq n$: The velocity of all the joints of the body skeleton.
- $A = (a_1, \ldots, a_i, \ldots, a_n)$ $1 \leq i \leq n$ The acceleration of all the joints of the body skeleton.
- $AA = (aa_1, \ldots, aa_i, \ldots, aa_n)$ $1 \leq i \leq n$ The angular acceleration of all the joints of the body skeleton.
- Quantity of Motion (QoM [137]): is a weighted average of the velocities of a set of representative joints in the body. It can be expressed by the following equation:

$$QoM = \Sigma_i v_i^2 \tag{5.22}$$

- Intensity of movement (IoM): which explains the suddenness or stability of a movement. It is calculated by:

$$IoM = \frac{1}{T}\Sigma_j a_i(t_j), \;\; 1 \leq j \leq T = The \;\; total \;\; number \;\; of \;\; frames. \tag{5.23}$$

Finally, our descriptor will be a matrix as follows.

$$MoF = \begin{pmatrix} Ea(t_1) & \Theta(t_1) & D(t_1) & V(t_1) & A(t_1) & AA(t_1) & QoM(t_1) & SC(t_1) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ Ea(t_T) & \Theta(t_T) & D(t_T) & V(t_T) & A(t_T) & AA(t_T) & QoM(t_T) & SC(t_T) \end{pmatrix} \tag{5.24}$$

Where $SC = [IoM \;\; C(j_h) \;\; T_{hh} \;\; T_{ht} \;\; T_{rhsh} \;\; T_{lhsh} \;\; q \;\; SE]$, and $t_j, 1 \leq j \leq T$ represent the number of frames of each gesture.

5.6 Expressive motion recognition and analysis using Feed Forward Neural Network

Artificial neural networks are a technique in machine learning inspired by biological samples of the brain and nervous system. The structure and function of the biological brain is quite different from that of a conventional digital computer. In many ways, the biological brain (the most complete and complex example of which is the human brain) is far more advanced and superior than conventional computers. The most important and unique advantage of the biological brain is its ability to learn and adapt or update. This is while a normal computer does not have such capabilities and features by itself. Ordinary computers basically perform certain tasks based on instructions stored on them, which we call them "software". The infrastructure and basic structure of neural networks are "neurons". In this field, neurons can be introduced as a processing unit. n a neural network, neurons typically communicate with each other through a "synaptic weight" or shortly "weight." In a neural network, neurons receive "weighted" information from the neurons to which they are attached through these synaptic connections, and by passing a weighted value through it, input signals (either external inputs from the environment or the outputs of other neurons) , produce the outputs. This process is called "activation function". Depending on the type of connection between the neurons, there are two main types of network architecture called "feed forward neural networks" and "recurrent neural networks". If there is no "feedback" from the neuron outputs to the input throughout the network, this network is called the "feed forward neural network." Otherwise, if there is such feedback, meaning that there is a synaptic connection from the outputs to the inputs (whether their inputs or the inputs of other neurons), then the network is called a "recurrent neural network". Neural networks are usually layered. Feed forward neural networks are divided into two categories, "single-layer" (Fig 5.25) or "multi-layer" (Fig 5.26), depending on the number of layers. Figure 5.25 shows a single layer fully connected feed forward neural network. There are two layers of input layer and output layer in this structure. Of course, it should be noted that the input layer is not calculated because no calculations are done in it. Weights are responsible for transmitting the signals of the input layer to the output layer, and the neurons in this layer are responsible for calculating the output signals. Figure 5.26) shows a fully connected multi layer feed-forward neural network with one “hidden layer”. Unlike a single-layer, a multi layer network has at least one hidden layers of neuron between the its input and output layers. According to [95], the function of neurons of hidden layers are to mediate between the input layer neurons and the output layer neurons. Having one or more hidden layers increases the network’s ability to extract higher-order statistics. As mentioned, pairs

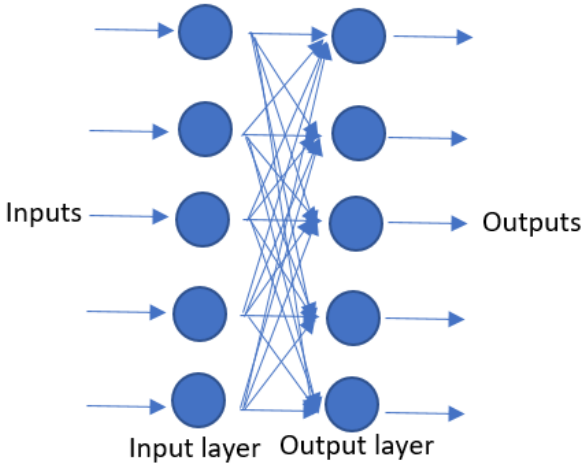


Figure 5.25: A single layer feed-forward neural network

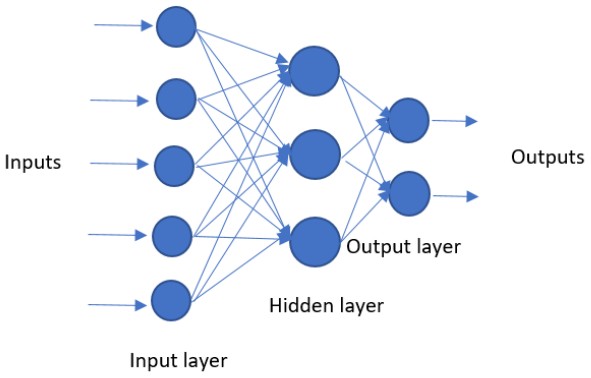


Figure 5.26: A multi-layer feed-forward neural network

of Figures 5.25 and 5.26 represent fully connected networks because each neuron in each layer is connected to all the neurons in its next layer. If some of the synaptic connections are missing, the network is called a "partial connected networks". Automatic learning is one of the unique features of the neural network that distinguishes it from a normal computer. This means that a neural network can learn from its environment and according to what it has learned, improve its performance and update its knowledge. Haykin [95], defined learning in the field of neural networks as follows:

"Learning is a process by which the free parameters of a neural network are adapted through a process of stimulation by the environment in which the network is embedded. The type of learning is determined by the manner in which the parameter changes take place [95]."

5.6.1 Back propagation algorithm

Among the many learning algorithms in the field, the "back propagation algorithm" is the most popular and most widely used for training through feed forward neural networks. In fact, this algorithm is a tool for updating the synaptic weights of networks by reproducing the gradient vector. In this vector, each element is defined as a derivative of the error measurement with respect to a parameter. To calculate these error signals, the difference between the actual network outputs and the desired outputs must be calculated. Therefore, in order to calculate it, one must have access to a set of desired outputs in order to enable the system to learn. For this reason, back propagation is called a supervised learning method. The following is a brief description of the back propagation algorithm for training a feed forward neural network. For this purpose, consider the multi layer neural network shown in Figure 5.26. We select a neuron in the output layer and call it neuron j. As mentioned in [95], the error signal at the output of neurons j for the n_{th} iteration is:

$$e_j(n) = d_j - y_j(n) \tag{5.25}$$

Where d_j is the desired output for neurons j and $y_j(n)$ is the real output of neurons j, which is calculated by the use of current weights of network in its n^{th} iteration. For a specific input, there is a specific output that the network is responsible for generating. Introduction each instance of the training data set is called as "iteration". The instantaneous amount of error energy for neuron j is defined as:

$$\epsilon_j(n) = \frac{1}{2} e_j^2(n) \tag{5.26}$$

Since the only neurons visible in neural networks are the cells present in the output layer, the error signals for these neurons can be calculated directly. Thus, the instantaneous value, (n) of the total error energy is equal to the sum of all j (n) calculated for all neurons in the output layer, as shown in:

$$\epsilon(n) = \frac{1}{2} \Sigma_{j \in Q} e_j^2(n) \tag{5.27}$$

Where Q is the set of all neurons in the output layer. For example, suppose there are N patterns in a training

data set. The average square energy for the network is as follow:

$$\epsilon_{av} = \frac{1}{N}\Sigma_{n=1}^{N}\epsilon(n) \tag{5.28}$$

It should also be noted that the instantaneous error energy (n) as well as the mean error energy, $_{av}$, are a function of all free parameters, which are the synaptic weights and the bias levels. The back propagation algorithm is, as will be explained, a method for tuning all free parameters of the neural network to minimize the mean error energy, $_{av}$. back propagation can be divided into "sequential mode mode" and "batch mode". In sequential mode, the weights are updated after presenting and entering a sample from the training set. A complete presentation of a training set is called "epoch". In batch mode, the weights are updated after all training samples are presented, i.e. after completing an epoch. Consecutive mode is also called on line, pattern or stochastic mode. This is the most common method of operation and is explained below. The output of a neuron j is calculated as follows:

$$y_j(n) = f(\Sigma_{i=0}^{m} w_{ji}(n)y_j(n)) \tag{5.29}$$

where m represents the total number of inputs to the neuron j (excluding the bias) from the previous layer and f is the activation function used in the neuron j, which can be a nonlinear function. Here w_{j0} equals the bias b_j applied to the neuron j and it corresponds to the fixed input $y_0 = +1$. Updates of the weights of neuron j are proportional to the partial derivatives of the instantaneous error energy (n) with respect to the corresponding weight, $\partial\epsilon(n)/\partial w_{ji}(n)$. Its calculation will be as follows:

$$\frac{\partial\epsilon(n)}{\partial w_{ji}(n)} = \frac{\partial\epsilon(n)}{\partial e_j(n)}\frac{\partial e_j(n)}{\partial y_j(n)}\frac{\partial y_j(n)}{\partial w_{ji}(n)} \tag{5.30}$$

From Equations 5.26, 5.25 and 5.29 respectively, Equation 5.31 is obtained.

$$\frac{\partial\epsilon(n)}{\partial e_j(n)} = e_j(n) \tag{5.31}$$

$$\frac{\partial e_j(n)}{\partial y_j(n)} = -1 \tag{5.32}$$

$$\frac{\partial y_j(n)}{\partial w_{ji}(n)} = f'(\Sigma_{i=1}^{m} w_{ji}(n))\frac{\partial(\Sigma_{i=0}^{m} w_{ji}(n)y_j(n))}{\partial w_{ji}(n)} = f'(\Sigma_{i=0}^{m} w_{ji}(n)y_j(n))y_j(n) \tag{5.33}$$

Where

$$f'(\Sigma_{i=0}^{m} w_{ji}(n)y_j(n)) = \frac{\partial f(\Sigma_{i=0}^{m} w_{ji}(n)y_j(n))}{\partial(\Sigma_{i=0}^{m} w_{ji}(n)y_j(n))} \tag{5.34}$$

Substituting Equations 5.31, 5.32 and 5.33 in Equation 5.30 yields Equation 5.35.

$$\frac{\partial\epsilon(n)}{\partial w_{ji}(n)} = -e_j(n) = f'(\Sigma_{i=0}^{m} w_{ji}(n)y_j(n)))y_j(n) \tag{5.35}$$

The correction $\Delta w_{ji}(n)$ applied to $w_j i(n)$ is defined by the delta rule, given in Equation 5.36.

$$\Delta w_{ji}(n) = -\eta \frac{\partial \epsilon(n)}{ji(n)} \tag{5.36}$$

Where η represents the learning-rate parameter of the back propagation algorithm. This parameter is usually set to a pre-defined value and will be kept constant during the whole of the algorithm. After the whole data set was trained by the feed forward neural network, we used ROC (Receiver Operating Characteristic)Curve and AUC to evaluate the whole process including the extracted feature as well as our network. An ROC curve consists of a graph which display the performance of a classification model at all classification thresholds. This curve shows two parameters:

- True Positive Rate.
- False positive Rate.

True Positive Rate (TPR) which can be another name for recall and is calculated as follows:

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{5.37}$$

Also False Positive Rate (FPR) is computed as follows:

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} \tag{5.38}$$

Where TP, FN, P, FP, N and TN are true positive, false negative, condition positive, condition negative, false positive and true negative respectively. In the confusion matrix, the accuracy of the correct classification is calculated. It is computed by considering the correct prediction ratio to the total number of predictions, as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.39}$$

5.6.2 Experimental results of the classification of expressive gestures with the feed forward neural network

In this section, the components described in the proposed descriptor are first calculated for all joints of the body parts. The input data of the artificial neural network presented in this work is an $m \times n$ matrix in which n is equal to the number of available examples and m is equal to the number of the components of the proposed descriptor. To this end, we calculated the mean, standard deviation and the range of all elements of our descriptor so that we could prepare the data for the input of this network (Fig 5.27). A summary of the components of the descriptor is shown in the table 5.5.

The next part which is identification phase is done in two section as follows:

Table 5.5: Descriptor components.

Geometric features	Euler angles. The angles around elbows (θ_l^1, θ_r^1), neck (θ_l^2, θ_r^2), hips (θ_l^3, θ_r^3) and knees (θ_l^4, θ_r^4). The distances between different parts of body: $D = (d_h, d_{hf}^l, d_{hf}^r, d_{hn}^r, d_{hn}^l, d_f)$. The curvature C. The triangle between two hands and the head, T_{hh}. The triangle between two hands and the torso, T_{ht}. The triangle between the right hand, the shoulder center and the head, T_{rhsh}. The triangle between the left hand, the shoulder center and the head, T_{lhsh}. The quadrilateral consists of two hands and feet, q. The minimum ellipse-shaped space occupied by the skeleton,SE.
Time depended features	The velocity of all the joints: $V = (v_1, \ldots, v_i, \ldots, v_n)\ \ 1 \leq i \leq n$. The acceleration of all the joints: $A = (a_1, \ldots, a_i, \ldots, a_n)\ \ 1 \leq i \leq n$. The angular acceleration of all the joints: $AA = (aa_1, \ldots, aa_i, \ldots, aa_n)\ \ 1 \leq i \leq n$. Quantity of Motion QoM. Intensity of movement IoM.

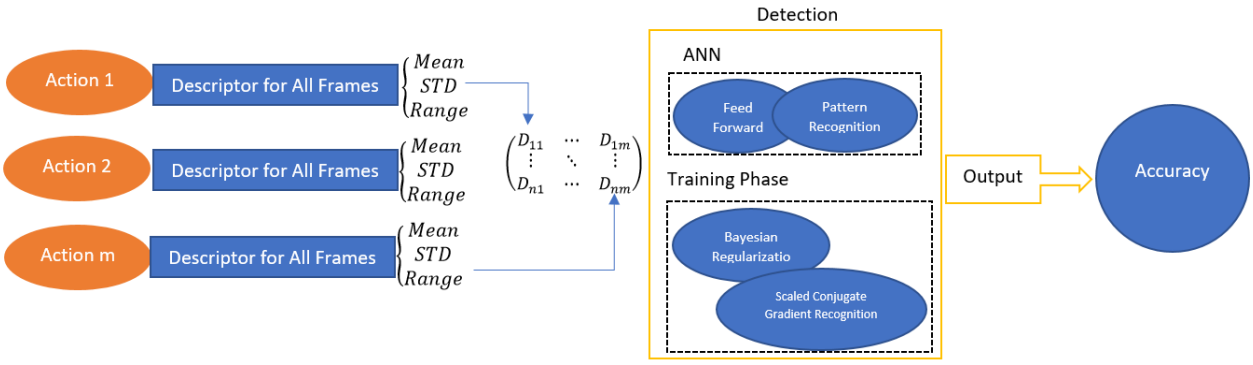


Figure 5.27: Proposed pipeline.

5.6.2.1 Gesture Recognition

The aim of this work is to analyze the performance of the proposed descriptor using Artificial Neural Network. All experiments are conducted in MATLAB 2018b. As mentioned, automatic recognition is done in two parts. In the first part, we tried to identify the movements performed in the data set using the proposed descriptor. So at this point, there are 5 classes to identify. To do this, in the first section of detection process, Feed forward and Pattern Recognition, 15 neurons were selected with a single hidden layer. 65% of the data were selected for training,10% for validation and 25% for test. In the final output layer of the described network, five neurons are used which belong to the classes as Dance, Move, Wave, Stop and Point, accordingly. The detection percentage in the first part was 100%. The confusion matrix and ROC diagram (threshold = 0.5) for action recognition are also shown in the Figure 5.28 . The number of epochs to reach the best validation performance, according to the cross entropy measurement is equal to 66. It is shown in Figure 5.29. After that, increasing the epochs will lead to decreasing the performance.

In this section, in order to prove the efficiency of the descriptor described in this work, we also implemented this descriptor in a public data-set called SYSU 3D HUMAN-OBJECT INTERACTION [102]. In this data set, 40 participants were asked to do 12 daily activities. In each of these gestures, participants interacted with six

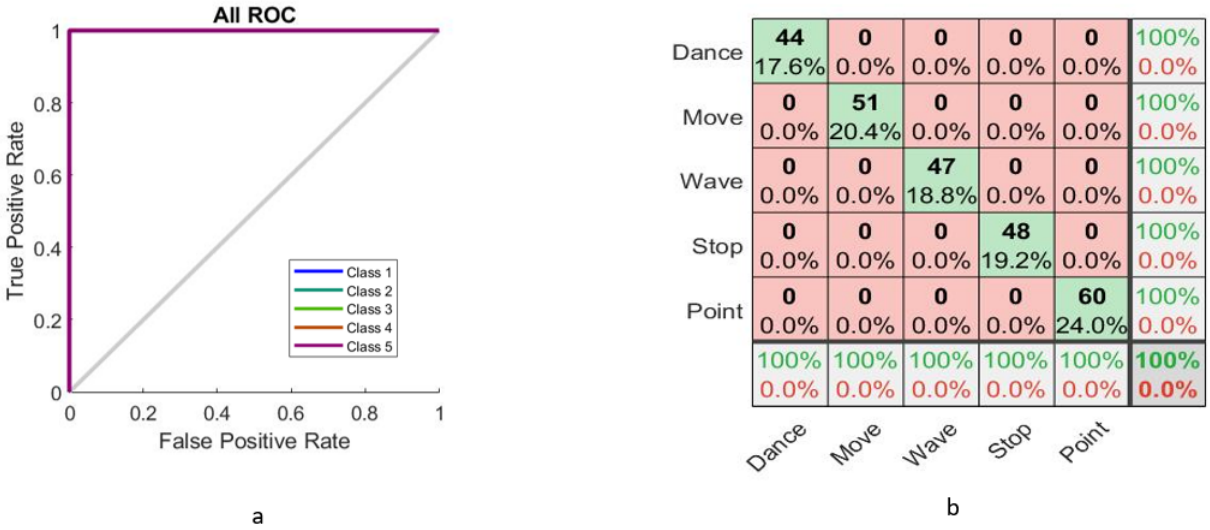


Figure 5.28: The ROC diagram and Confusion Matrix for Action Recognition

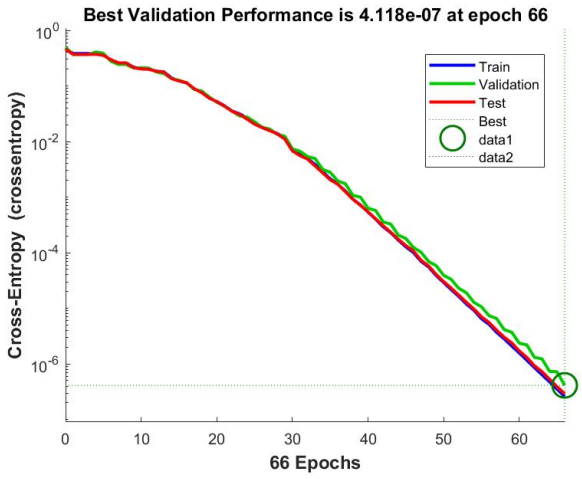


Figure 5.29: The Performance Diagram

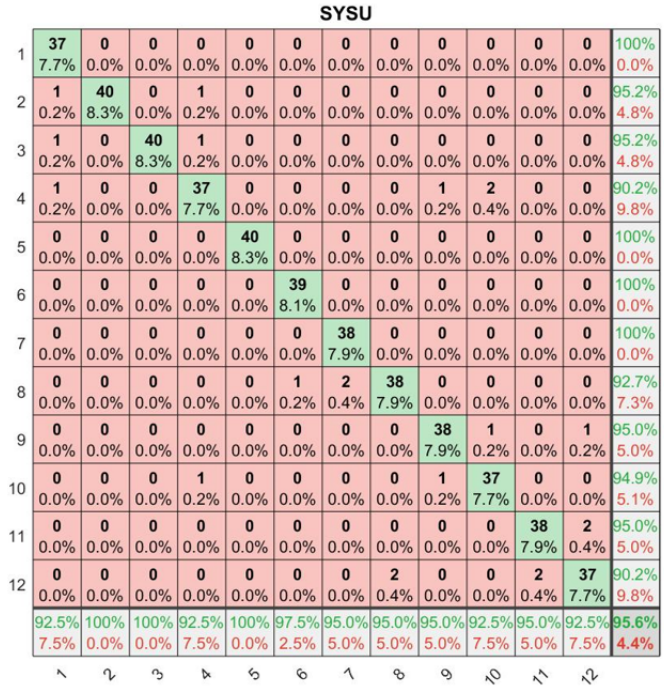


Figure 5.30: Confusion matrix of SYSU HOI

Method	Year	Accuracy(%)
Laban Movement Analysis and Dynamic Time Warping [199]	2020	92.32
Part-Level Graph Convolutional Network [147]	2020	84.2
Convolutional Neural Networks with Joint Supervision [104]	2019	97.08
Geometric Algebra Representation and Ensemble Action Classification [49]	2019	84.62
Proposed method	2020	**95.6**

Table 5.6: Comparison with the state-of-the-art results SYSU HOI data-set.

different objects: phone, chair, bag, wallet, mop and besom. So there are a total of 480 sequences in this data set, each sequence using the Kinect sensor gives the information about the RGB frames, depth sequence and 3D coordinates of skeleton data. Using this descriptor, as well as the feed forward neural network algorithm, the accuracy percentage is 95.6%.The relevant confusion matrix is shown in the Figure 5.33. The table 5.6 compares the results of implementing the proposed model on the SYSU HOI data set with other available results in the literature. As can be seen, the result obtained from this method is in many cases outperforms the other methods and the detection percentage is higher, so it can be concluded that this descriptor is a strong one.

5.6.2.2 Emotion Recognition

In the second part of this section, we will identify the emotions that are expressed by the five movements mentioned above. For this purpose, we first divide the data into five parts. Therefore, the input data to the network has 200 samples. And the output of the network has 4 samples, called Anger, Neutral, Happiness and Sadness. Because the number of samples is lower in the emotion recognition section, more data should be devoted to training. For example for "Dance" gesture, we use 80% of the data for training, 5% for validation, and 15% for testing. The results obtained from 15 neurons in the hidden layer for the dance gesture are shown in Figure 5.31.

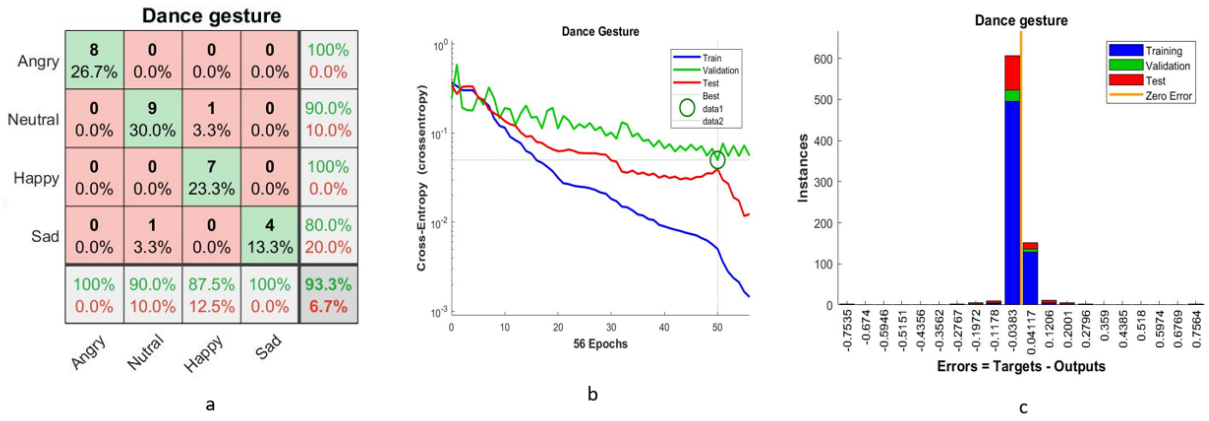


Figure 5.31: a: Confusion matrix, b: Performance validation and c: Histogram error for dance gesture

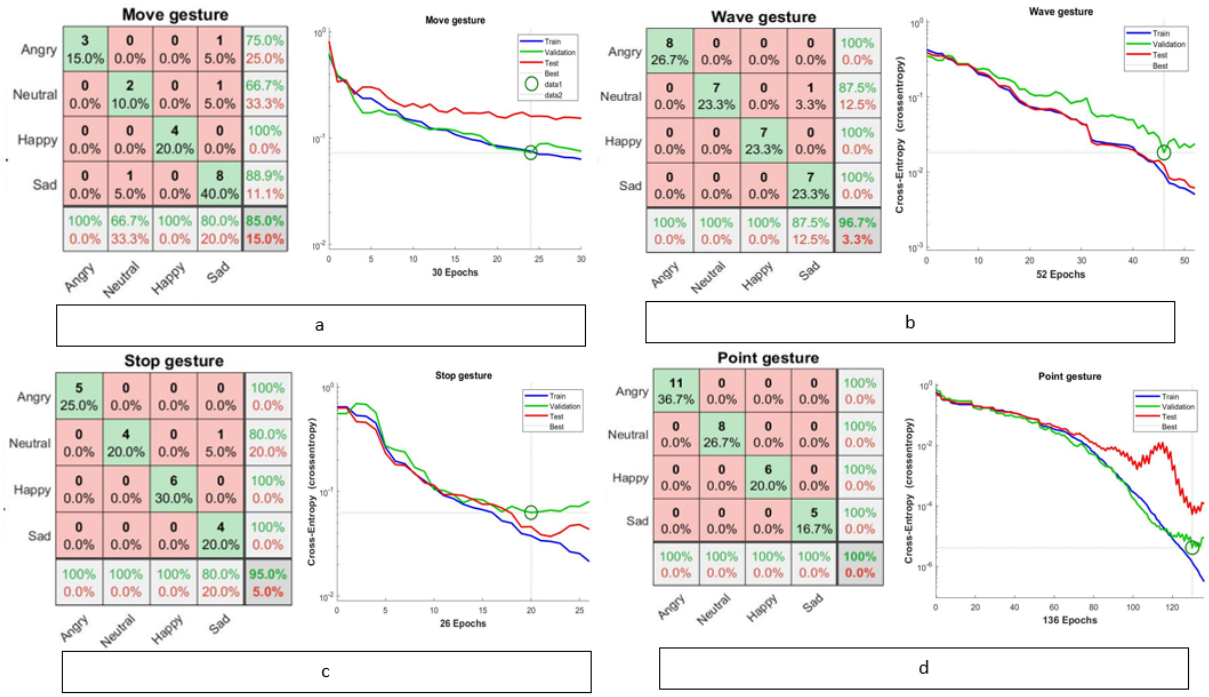


Figure 5.32: Confusion matrix and Performance validation for: a) Move, b) Wave, c) Stop and d) Point gestures

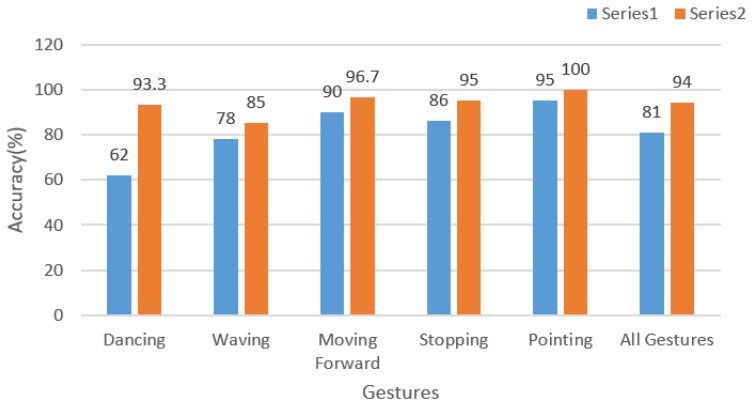


Figure 5.33: Comparison of the results obtained in [7], Series1, and proposed method, Series2.

The confusion matrices and performance validation diagrams for neutral, happy, and sad gesture, respectively, are shown in Figure 5.32. For the "move" gesture, the best results were obtained by allocating 85% of the data for training, 5% for validation, and 10% for testing, and 16 neurons in the hidden layer. For the "Wave" gesture, the best results were obtained by allocating 70% of the data for training, 10% for validation, and 20% for testing, and 12 neurons in the hidden layer. For the "Wave" gesture, the best results were obtained by allocating 80% of the data for training, 10% for validation, and 10% for testing, and 20 neurons in the hidden layer. For the "Point" gesture, the best results were obtained by allocating 80% of the data for training, 5% for validation, and 15% for testing, and 10 neurons in the hidden layer. In all cases, the data is divided randomly. The chart 5.33 shows the difference between the results of this method and the results of [7]. All comparisons in this dissertation are based on the pipeline (combination of descriptor and used method).

5.6.3 Selection of relevant characteristics with the step wise regression method.

One of the main and important problems in decision regression analysis is how to select the variables that have the greatest impact on pattern building. Researchers often use a number of predictor variables in research. The large number of these variables increases the complexity of the calculations and consequently the cost of making the template. For this reason, they are trying to create a suitable model with a subset relatively smaller than the original set. During the process of building a model, researchers can test many methods and algorithms and even combine several methods based on theories. However, it should be noted that this whole process may take a considerable amount of time and increase research costs. One of the most common methods in selecting variables is the step-by-step method, which is often used in articles that are based on experience [105]. The purpose of the stepwise regression algorithm is to select a model step by step. In this process, adding or removing a predictor occurs only on the basis of statistical significance. The result of this process is a single regression model. Stepwise analysis has two methods, forward or backward. Forward progress has been used more than backward analysis. Today, thanks to software such as Minitab, researchers can monitor and control the details of the process, including the level of importance and manipulation of variables (such as adding or removing a predictor).

In this regards, we use the step wise regression method to define a subset of the descriptors described in the

section above that have the greatest impact on emotion recognition. Consider the following regression model

$$Y = Xb + \epsilon \tag{5.40}$$

Where Y is an q-vector of labels and $X = (x_{i,j}) = (X_1, X_2, \ldots, X_p)$ is an $T \times P$ matrix of feature, b is an $p-$vector of coefficient estimates from step wise regression, $\epsilon = (\epsilon_1, \ldots, \epsilon_q) \sim N(O, \sigma^2 I_n)$, where $\sigma^2 \geq 1$ is an unknown but fixed constant and I_q denotes the $q \times q$ identity matrix. $X_1, X_2, \ldots, X_p$ is used to represent the column of X. This method uses forward selection and backward elimination methods to calculate the output vector, which determines the coefficient of importance of features. The algorithm starts by calculating all bi-variate r^2 parameters for all dependent and independent variables. Afterward, the independent variable with largest r^2 will be selected. In the next step, the remaining independent variables are added separately to the model with the "best" independent variable, and stepwise selects the independent variable that causes the largest increase in R^2. Next, the stepwise regression computes the association of the rest parameters in the model with the selected best variable and the second "biggest contribution" variable will be selected. In the same way, the variable that can cause the largest increase in R^2 with the two predictors selected in the previous steps is selected as the third predictor. The forward selection method is stopped in conventional statistical packages if the increase of R^2 from one stage versus R^2 in the previous stage is not statistically significant. It is important that, over the process of selecting parameters, stepwise process utilizes the equation ($F_{calculated} = [(R^2_{Larger} - R^2_{Small})/(k_L - k_S)]/[(1 - R^2_{Larger})/(n - k_L - 1)]$) to evaluate the null hypothesis $H_0 : R^2_{Larger} = R^2_{Small}$. Specifically, k_L is the number of predictors used to obtain R^2_{Larger}, and k_S is the number of predictors used to obtain $R^2_{Smaller}$. The degrees of freedom are $(k_L - k_S)$for the numerator, and $(n - k_L - 1)$ for the denominator [247]. Then the backward phase begins with all the remained features. The deletion of each feature (if any) is evaluated using a chosen model fit criterion. It eliminates the feature whose loss gives the most statistically negligible decay of the model fit, and repeats this process until no more feature can be deleted without a statistically negligible loss of fit. Computationally, this algorithm uses the partial F-statistic, F, in order to select variables in its process which is defined as follow:

$$F = \frac{MSR}{MSE} \tag{5.41}$$

Which MSR is the mean square due to regression and the MSE is the mean square of the error for selecting relevant variables. The variable x with the highest values of F is considered as the candidate variable to be added to the model. Therefore, if the value of F is higher than a predetermined level, it is selected. So the variable X is added, otherwise the program ends and no X variable is useful enough to enter the regression [173].

5.6.4 Results of relevant characteristics with stepwise regression.

In the last part of this chapter, we examine the effect of all components of the proposed descriptor on the emotions classification. For this purpose, for each of the 5 gestures in this XEM data set, we examined all the components using a step wise regression function. As it is mentioned step-wise regression involves two methods of forward and backward propagation, which are repeated alternately. Step-wise regression is the change in forward selection, so that after each iteration a variable is added, these added variables are nominated to be

Table 5.7: The most important elements for recognizing emotions in each gesture.

Dancing	STD(d^r_{hf}, $V_{RightHand}$, $V_{LeftHand}$, $A_{RightHand}$, $A_{LeftFoot}$, $A_{HipCenter}$, $A_{LeftHand}$, $EA_{HipCenter}$, $EA_{LeftShoulder}$); Range(θ^2_r, θ^4_l, A_{Elbow}, $A_{RightFoot}$, $AA_{RightHand}$, $EA_{RightShoulder}$, $EA_{RightElbow}$, $EA_{LeftShoulder}$); Mean(θ^1_l, θ^1_r, θ^3_r, θ^4_l, $AA_{SholuderCenter}$, $AA_{RightHand}$, $AA_{LeftHip}$, EA_{Neck}, $EA_{RightSHoulder}$, EA_{Elbow}, $EA_{RightHand}$, $EA_{LeftFoot}$, T_{rhsh}); $C_{RightHand}$
Moving	STD(d_h, θ^r_2, θ^l_2, θ^l_1, θ^r_1, θ^l_3, $V_{HipCenter}$, $Area_q$); Range(d^l_{hn}, θ^l_4, θ^r_4, $V_{HipCenter}$, $A_{HipCenter}$, $EA_{RightHand}$, $EA_{LeftHand}$, $EA_{RightElbow}$, $EA_{LeftElbow}$); Mean(d_h, d^r_{hn}, d_{ff}, θ^r_1, $V_{RightHand}$, $V_{LeftHand}$, $V_{RightHip}$, $V_{LeftHip}$, $EA_{RightHand}$, $EA_{LeftHand}$, $EA_{RightElbow}$, $EA_{LeftElbow}$, $Area(T_{hh})$, $Area(T_{rhsh})$, $Area(T_{lhsh})$, $Area_q$, $Volum_{SE}$); $IOM_{LeftHand}$; $IOM_{RightHand}$
Waving	STD, Range, Mean (d^l_{hf}, d^r_{hn}, d^r_{hh}, θ^r_1, θ^l_1, $V_{RightHand}$, $V_{LeftHand}$); Mean, Range ($Area(T_{hh,lhsh,rhsh,ht,q})$); $IOM_{LeftHand}$; $IOM_{RightHand}$; $C_{RightHand}$; $C_{LeftHand}$
Stopping	STD (d^r_{hn}, θ^l_3, V_{Neck}, $AA_{RightShoulder}$, $AA_{LeftShoulder}$, $AA_{RightKnee}$, $EA_{HipCenter}$, $Area_q$); Mean(d^r_{hf}, $A_{RightHand}$, $AA_{RightHand}$, AA_{Head}, $EA_{LeftKnee}$, $Volum_{SE}$); Range(d_{hh}, θ^l_2, $A_{HipCenter}$, $AA_{HipCenter}$, $Area(T_{rsh})$, $Area(T_{lsh})$, $Area(T_{hh})$, $Volum_{SE}$); $C_{RightHand}$; $C_{LeftHand}$
Pointing	STD(θ^r_3, θ^l_3, V_{Neck}, $V_{RightHand}$, $AA_{RightElbow}$); Mean(d^{hf}, θ^l_1, θ^r_3, θ^l_3, V_{Neck}, $V_{RightHand}$, $A_{RightHand}$, $AA_{RightElbow}$), $EA_{RightElbow}$), $Area_{rhsh}$, $Volum_{SE}$); Range(d^r_{hf}, θ^l_1, θ^r_3, θ^l_3, $Volum_{SE}$)); $C_{RightHand}$; $IOM_{RightHand}$; $QOM_{RightHand}$

examined in the model to see how important they are in the identification stage. If a variable with a small effect is found, it exits the model. Therefore, it requires two important steps: one to put the variable in the model and the other to remove the variable. In order to avoid the infinite repetition of the algorithm, we must note that the threshold probability of adding a variable must be less than the threshold probability of its deletion. By calculating this probability, the factors that are most effective in identifying emotions in each gesture are shown in the table 5.7.

As can be seen in the table above, there is a symmetry between these factors, for example when the speed of the right hand is important in the "dance" gesture, the speed of the left hand is also important. The next thing to notice is that in complex movements where all the parts of the body move, the number of important factors is less, while in simple movements, for example, "pointing", in which only the hand moves, many factors are involved to identify emotions. Also, time-dependent factors are very important in identifying emotions. And that's why the velocity and acceleration of the part of the body that move is an important factor in all gestures. Therefore, according to these results, we can decide that sadness can be detected with low speed and low acceleration and low force, happiness can be detected with high speed and soft movements and anger with high speed and high force.

5.7 Conclusion

The aim of this study is to characterize and recognize human emotions expressed through body movement first based on the machine learning method and then on the human approach to assess the performance of our

recognition system. We obtained the following conclusions:

- In the study of emotion recognition, the human and RDF classifiers were able to distinguish between the 4 states, with some confusion between the emotions of joy and anger and between the two neutral and sad states. If we compare the two results, we find that the learning method was more precise than the observers. This may be due to the limited number of observers who participated in the evaluation of the different sequences. In addition, this can be explained by the type of gesture chosen. We try to recognize emotions through limited gestures. For example, by performing the pointing gesture with the emotion of sadness or with the neutral state, we will have almost the same movement with a somewhat stable rhythm. So, visually, it will be difficult for observers to distinguish the two types of movement.

- In studying the importance of characteristics with the human approach, each observer assessed the importance of each LMA factor in the characterization of each emotion. With the RDF classifier, the algorithm consists in studying the importance of the characteristics in the characterization of the emotions compared to the neutral state. Additionally, we apply Turkey's test to remove redundant features. As presented in Table5.3, the emotion of sadness was characterized by the factors of the two Stress-Fitness components, with the exception of the directional movement factor. The same result is obtained in the evaluation of observers in Table 5.4.However, for the emotions of joy and anger, the RDF classifier found that the two most discriminating characteristics in characterizing these two emotions are the time and weight factors. These results confirm that our proposed method makes it possible to characterize emotions and define important characteristics while optimizing our motion descriptor by keeping only the characteristics both relevant and non-redundant.

- In the next step, we tried to increase the accuracy of our work by changing the descriptor, for this purpose, we created another descriptor using geometric factors and time-dependent factors. In defining this descriptor, two points have been considered. The first point was to be able to detect movements. For this reason, factors such as distances, angles and other geometric factors are included in this descriptor. The second point is the descriptor's second task in identifying emotion. And to identify the emotions, we chose the factors that were very expressive when recording the data set. For example, we found that when participants were told to imagine that they were sad when they were moving, they would move slowly, make soft movements and their body shrink. Or when they were angry, they did the movements with high speed and high force. And that's why we've chosen factors like speed, acceleration, Qom and IoM. In order to evaluate this descriptor, we used the feed forward neural network method. And we did that assessment in two steps. In the first step, we used this descriptor to identify the movements. Identification of gestures was performed in two data set, XEM and SYSU HOI data-set. In the XEM, the percentage of gesture detection was 100%. Because this is a small data set that has a small number of gestures, and on the other hand the gestures are completely different, such accuracy is expected. And that was one reason why we should consider another public data set. Using the algorithm presented in this method, the detection accuracy in SYSU data set was 95.6%. This percentage was higher than many studies in the literature. In the next part of the identification phase, we identified the emotions which the gestures were performed with. Identifying emotions in the gesture of "moving" had the least accuracy of

recognition. The reason could be that only two hands were used in this movement and the other parts of the body had no effect on this gesture. While, as shown in the table 5.7, the angles of the neck, the distances between the legs, and the volume of space used by the body are important factors, all of which were almost constant in this gesture. The improvement of the results using the method proposed in this work is specified in the diagram 5.33. In the last part of this chapter, we examined the important factors for identifying emotions in each gesture. Using the results, we found that speed, strength, contraction and expansion of body parts, softness and firmness of movements are among the factors that are important for identifying emotions.

Note: Since the size of XEM dataset is small, it was not possible to recognize emotions without considering the gesture and its recognition accuracy was very low.

Chapter 6

General conclusion and perspectives

Contents

6.1 Conclusion

In this work, we deal with the problem of gesture recognition by developing a robust and efficient system which considers the gesture and also the emotion of the person. Three approaches are carried out in this work:

The first one is to recognize the dynamic gestures of human by the use of dynamic time warping algorithm. A local motion descriptor is implemented for the representation of motion. A spline function is used in order to minimize the descriptor size to the inputs of the DTW model. A contribution is made to the Support Vector Machine model to improve recognition rates under conditions of similarity between movements by the use of DTW. We made a new configuration on the SYSU HOI data set, which allowed us to increase the detection accuracy by 5.69%. The evaluation of the system is made on four public data set. The recognition rates obtained vary between 86.63% and 100%.

- The second approach was to develop a system for recognizing expressive gestures with global learning methods. A global descriptor is created to describe the entire movement and its expressiveness. An adjustment of the different parameters of the learning methods is carried out followed by a comparative study between these methods in order to select the best one. The evaluation of the system is made on a public basis and our basis is made up of expressive gestures. This approach allowed us to choose the Feed Forward Neural Network method for the best method.

- The third approach is to evaluate our established system with reference to the human approach. This is a statistical study which is based on the opinions of a set of observers in the perception of emotions and also the evaluation of the proposed movement descriptor. A feature selection algorithm is set up to study the importance of each for the expression of each emotion. Finally, the results from the machine learning method are compared with those from the human approach to conclude on the reliability of our system.

Our gesture recognition system succeeded in classifying emotions like a human and selecting common relevant characteristics with those chosen by the human approach while optimizing the size of our motion descriptor.

Our system processes data with low computational complexity thanks to the proposed algorithms (sampling, normalizing, selection of characteristics) and achieves good prediction accuracy. Our results are comparable to specialized state-of-the-art results.

6.2 Perspectives

As perspectives, it would be interesting to explore our gesture recognition system in a robotic application like ours which consists in controlling the NAO robot via gestures. By applying our system, the robot will be able to recognize the person's gestures and also his emotion through his movement. Thus, it can do the tasks associated with gestures while interacting with the person according to his mood, which makes the interaction between the two parts more natural. However, this goal requires some improvements in our system:

- In the gesture recognition part, our system works in "offline" mode. So we plan to make the operation in "online" mode. This requires the phase of the detection of the beginning and end of the gesture.
- In the emotional part, some perspectives are considered such as: improving our expressive data set to recognize a greater number of emotions. Another idea for expanding this data set is to include movements that are not pre-defined. Also, daily movements such as knocking, walking, cleaning the house, answering the phone and etc. will be added to this collection. Other modalities such as facial expressions, audio, etc. can be combined with these gestures to increase the accuracy of identification and establishing a natural interaction.
- Study of the impact of other factors on the expression and recognition of emotions (gender, age, culture, etc.).
- In the field of bi-lateral interaction, we first intend to train NAO to perform a gesture with different emotions. And in the next step, the second NAO recognizes the feelings of the first one.

Appendix A

Support Vector Machine

In general, machine learning is done in both supervised and unsupervised ways. Many supervised machine learning methods work by giving a set of input vectors such as $X = \{x_n\}$ and their corresponding output vectors$T = \{t_n\}$ [31]. The goal is for the machine to be able to predict t using this training data for the new x input. In this regard, two distinct modes can be considered: Regression, in which t is a continuous variable, and Classification, in which t belongs to a discrete set. They are in the second category and the goal is to train the machine in such a way that the machine can do this classification well. In the learning process, the system needs to be trained first and then tested for new input values. Mathematically, the machine learning problem can be thought of as a mapping in which $X_i \longrightarrow Y_i$. In fact, a machine is defined by a set of possible mappings as $X \longrightarrow f(X, \alpha)$ in which the functions $f(X, \alpha)$ themselves can be adjusted by the tag. It is assumed that the system is deterministic and that the value of a particular input x and the choice of α always give a specific output equal to $f(X, \alpha)$. Choosing the right α is the same thing a trained machine does. Here we will focus on the case where the prediction $y(x, w)$ is expressed by the linear combination of the base function $\Phi_m(X)$ in the following form:

$$\Sigma_{m=0}^{m} w_m \Phi_m(X) = W^T \Phi \tag{A.1}$$

Where m_ws are model parameters called weights. In SVM, basic functions are used as Kernel functions, which for every m_x in the instruction set we have $\Phi_m(X) = K(X, X_m)$, where $K(.,.)$ is the Kernel function. Weight estimation in SVM is achieved by optimizing criteria that simultaneously try to minimize the $y(x, w)$ function. As a result, a number of weights are zeroed, resulting in a Sparse Model in which the prediction is managed by Equation A.1 and depends only on a subset of the Kernel function. In recent years, the use of support vector machines (SVM) has received much attention. It has been experimentally shown that the use of SVM in applications such as handwriting recognition, face recognition, etc., has achieved good results. Until its presentation in [139], the use of support vector machines was limited to a specific group of researchers. The reason for this limitation in the use of SVM is the slowness of the training algorithm for this method, especially in the case of large training sets. In other words, the SVM training algorithm has been complex and difficult for many engineers, and therefore new solutions and methods have been proposed to deal with this problem.

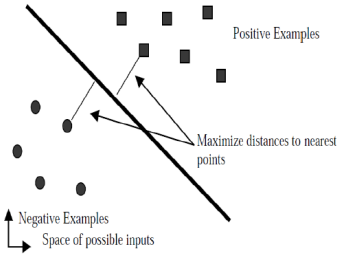


Figure A.1: SVM as a super plan for linear separation of samples in a data set.

A.1 Overview of SVM

In 1919, Support Vector Machines were introduced by Vapnik [256]. In its simplest form, linear SVM, a SVM is a super plan that separates a set of positive and negative samples with a maximum margin (Fig A.1 . The issue of data classification by SVM can be examined in several different ways:

- Linear SVM while having two type of data:
  - The data can be divided into two categories, Separable Data.
  - The data of the two categories are not separable into two separate categories, Nonseparable Data.
- Nonlinear SVM.
- Multi class SVM.

A.1.1 Separable Data

The simplest case is one in which the machine is trained linearly on the separable data. It is interesting that the equations obtained from this case can be generalized to the nonlinear and inseparable case, so this case is the basis for defining other cases. It should be noted that in this case it is assumed that the data are exactly in two categories. Note that the equations obtained from this case can be generalized to nonlinear and inseparable case, so this case is the basis for defining other cases. It should be noted that in this situation it is assumed that the data fall into exactly two categories. In this case, it is assumed that there is a set of separable training examples that can be labeled with y_i. In this case, the samples are expressed as regular pairs (x_i, y_i) in which $i = 1,\ldots,l$and $x_i \in \mathbb{R}$ and $y_i \in \{-1,1\}$. As shown in Figure A.2, the data can be separated and categorized by several separator lines (or superplanes in n-dimensional space). "Margin" is defined as the distance of the superscript to the nearest negative and positive instance.Each separating line in 2D space is written as the equation of the following line [191]:

$$w_1 x_1 + w_2 x_2 + b = 0 \tag{A.2}$$

And in 3D space will be as follows:

$$w_1 x_1 + w_2 x_2 + + w_3 x_3 + b = 0 \tag{A.3}$$

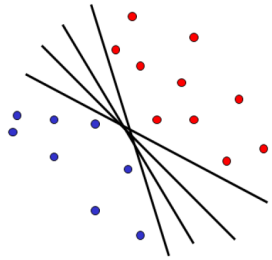


Figure A.2: Separation of data in space by different superplanes.

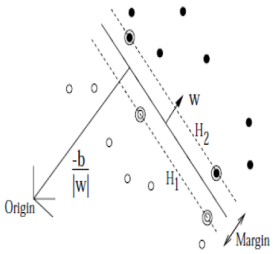


Figure A.3: Two-dimensional state assuming the bias value is negative.

In general, for a separating hyper plane, we have:

$$\Sigma_i w_i x_i + b = 0 \tag{A.4}$$

Which can be expressed in the following way:

$$u = \vec{w}.\vec{x} + b \tag{A.5}$$

Where w is the weight vector perpendicular to the super plane and b is the bias value. In this view, $u = 0$ refers to the separator super plane, and the nearest points are on the page $u = \pm 1$. In fact, assuming the two classes of data are positive and negative, the boundary vectors will be placed on the following super planes:

$$\vec{w}.\vec{x} + b = \pm 1 \tag{A.6}$$

The area between these two super planes is called the margin. Figure A.3 shows the two-dimensional state assuming the bias value is negative: As can be seen in the figure A.3, the space is divided into two groups of samples with the following properties:

$$\vec{w}.\vec{x} + b \geq 1 \qquad for \qquad y_i = 1 \tag{A.7}$$

And

$$\vec{w}.\vec{x} + b \leq -1 \qquad for \qquad y_i = -1 \tag{A.8}$$

The above formula can be combined as follows:

$$y_i(\vec{w}.\vec{x}) + b \geq 0 \qquad \forall i \tag{A.9}$$

In this case, the distance to the origin vertically (the closest distance) for points on the cloud $x.w + b = +1$ is equal to:

$$\frac{|1-b|}{\|W\|} \tag{A.10}$$

And similarly the distance to the origin vertically for the points on the plane cloud $x.w + b = -1$ is equal to [40]:

$$\frac{|-1-b|}{\|W\|} \tag{A.11}$$

And the distance from the source to the separator super plane is equal to:

$$\frac{|b|}{\|W\|} \tag{A.12}$$

Therefore, the minimum distance between this hyper plane to each of the mentioned pages will be as follows:

$$d_+ = d_- = \frac{1}{\|W\|} \tag{A.13}$$

Therefore, the mentioned margin, ie the distance between the two desired page clouds, is obtained as follows:

$$d_+ + d_- = \frac{2}{\|W\|} \tag{A.14}$$

In this way, the maximum margin can be expressed as the following constrained optimization equation:

$$\min \frac{1}{2}(\|w\|)^2 \qquad Subjectto : y_i(\vec{w}.\vec{x} + b) - 1 \geq 0 \qquad \forall i \tag{A.15}$$

To solve this problem, it is easier to raise and solve the problem in a dual way. To obtain the dual form of the problem, the positive Lagrangian coefficients $\alpha_i \geq 0$ are multiplied by the constraints and subtracted from the objective function, resulting in the following equation, which is called the Primal Problem:

$$L_p = \frac{1}{2}(\|w\|)^2 - \Sigma_i \alpha_i (y_i(\vec{w}.\vec{x} + b) - 1) \tag{A.16}$$

The Karush-Kuhn-Tucker (KKT) conditions play an important role in constrained optimization problems. These conditions represent the necessary and sufficient conditions to have the optimal response for the constrained equations and the derivative of the function with respect to the variables must be zero. By applying the KKT condition to L_p and if it is derived from L_p with respect to w and b and set to zero, we have:

$$w = \Sigma_i \alpha_i y_i x_i \tag{A.17}$$

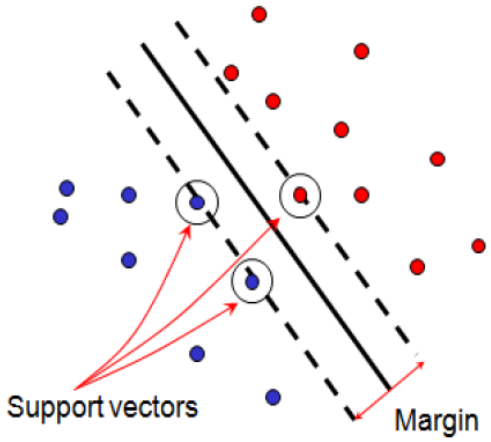


Figure A.4: Support Vectors.

$$0 = \Sigma_i \alpha_i y_i \tag{A.18}$$

In fact, the Primal Problem for (l_p), the KKT terms are as follows:

$$\frac{\partial}{\partial w_v} L_p = w_v - \Sigma \alpha_i y_i x_{iv} = 0 \qquad v = 1, \ldots d. \tag{A.19}$$

$$\frac{\partial}{\partial b} L_p = -\Sigma \alpha_i y_i = 0 \tag{A.20}$$

$$y_i(\vec{w}.\vec{x} + b) - 1 \geq 0 \qquad i = 1, \ldots L \qquad \alpha_i \geq 0 \qquad \forall i \tag{A.21}$$

$$\alpha_i(y_i(\vec{w}.\vec{x} + b) - 1) \geq 0 \qquad \forall i \tag{A.22}$$

The result is used to obtain the dual problem (DL). The KKT condition satisfies any constrained optimization problem, whether convex or non-convex (with any type of constraint. The SVM problem is a convex problem, and for convex problems, KKT conditions are necessary and sufficient conditions for w, b, α to be the solution to the problem, so solving the SVM problem is equivalent to finding a solution to the KKT condition, which has been used in various ways to find the SVM solution. The application of this condition is that if w is specified in the training procedure and b is not specified, b can be obtained.By replacing in Equation A.16 we will have:

$$L_D = \Sigma_i \alpha_i - \frac{1}{2} \Sigma_i \Sigma_j \alpha_i \alpha_j y_i y_j (\vec{x_i}.\vec{x_j}) \tag{A.23}$$

Which is called the Dual Problem. L_p and L_D are both obtained from the same condition, so by calculating the minimum of L_p or the maximum of L_D (which is the dual of L_p) with the condition $\alpha_i \geq 0$ the answer will be find. There is a Lagrangian coefficient $\alpha_i \geq 0$ for each training sample (x_i). In solving the problem, the samples for which $\alpha_i > 0$ are called support vectors, which are placed on the super planes of Equation 5.30 (Fig A.4).

A.1.2 Non-Linearly separable data

In this case, it is assumed that the existing data are not easily separable into two categories. In fact, in this case, which is closer to reality, the data is accompanied by noise and it is necessary to generalize the SVM to overcome this case. This situation can be solved by defining a soft margin [44], because if the above algorithm is applied to a set of inseparable elements for the separable state, no possible solution will be found. To extend the idea presented in the previous section to this state, a variable is defined as a positive Slack variable in the constraints of the problem, and we will have:

$$\vec{w}.\vec{x}+b \geq 1-\epsilon_i \qquad for \qquad y_i = 1 \tag{A.24}$$

$$\vec{w}.\vec{x}+b \geq -1+\epsilon_i \qquad for \qquad y_i = -1 \tag{A.25}$$

$$\epsilon_i \geq 0 \forall i \tag{A.26}$$

Therefore, for each error that occurs, each ϵ_i must be incremented separately. Therefore, $\Sigma_i \epsilon_i$ is a high limit for the number of training errors. So the best way to attribute an additional cost to errors is to change the objective function to minimize the function $\frac{(\|w\|)^2}{2} + C(\Sigma_i \epsilon_i)^2$ instead of $\frac{(\|w\|)^2}{2}$. The value of C is selected by the user. A large C value means more fines for errors. The proposed function is a convex function for any positive value of k, especially for $k = 1$ and $k = 2$ becomes a quadratic programming problem. By calculating the duality of the same function with the separable state, we get to:

$$Maximize: L_D = \Sigma_i \alpha_i - \frac{1}{2}\Sigma_i \Sigma_j \alpha_j y_i y_j (\vec{x_i}.\vec{x_j}) \tag{A.27}$$

$$Subject \ \ to: 0 \leq \alpha_i \leq C, \qquad \Sigma_i \alpha_i y_i = 0 \tag{A.28}$$

So we will have:

$$w = \Sigma_{i=1}^{N_s} \alpha_i y_i x_i \tag{A.29}$$

Where N_s refers to the number of support vectors. Therefore, the only difference with the optimal super plane state is that in this case α_i has C as a boundary above [40]. And the Lagrangian primal problem will be:

$$L_p = \frac{(\|w\|)^2}{2} + C(\Sigma_i \epsilon_i)^k - \Sigma_i \alpha_i \{y_i(x_i.w+b) - 1 + \epsilon_i\} - \Sigma \mu_i \epsilon_i \tag{A.30}$$

Where μ_i are Lagrange coefficients that cause positive ϵ_i. By applying KKT conditions to the resulting primal problem (L_p) we will have:

$$\frac{\partial}{v} L_p = w_v - \Sigma_i \alpha_i y_i x_{iv} = 0, \qquad v = 1, \ldots, d. \tag{A.31}$$

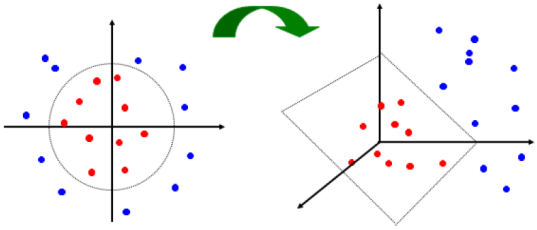


Figure A.5: Transfer the data from two-dimensional to three-dimensional space.

$$\frac{\partial}{\partial b} L_p = -\Sigma_i \alpha_i y_i = 0 \tag{A.32}$$

$$\frac{\partial}{\partial \epsilon_i} L_p = C - \alpha_i - \mu_i = 0 \tag{A.33}$$

By applying these conditions in L_p, LD (Eq A.27) can be reached. After calculating the Lagrangian coefficients, $0\alpha_i C$ are the supporting vectors for this state.

A.1.3 Nonlinear support vector machines

In a case where the data are not easily separated, a linear separator cannot be effective. But if we move the data to a larger space, a solution can be found to separate them. For example, Figure A.5 shows an example of this transition. As can be seen, after this transfer, the classification is applicable. In the cases mentioned in the previous sections, in fact, the internal product of the input training data in the support vectors (A.27) has been used to form a linear separator in the form of a super plane. Here we first map the data to an Euclidean space with higher dimensions and then use the internal multiplication of the obtained elements [40].

$$\Phi : \mathbb{R}^d \rightarrow \mathbb{H} \tag{A.34}$$

This vector space with higher dimension is called the Hilbert space [290]. Therefore, the training algorithm related to this internal multiplication in space H will be as $\Phi(x_i).\Phi(x_j)$. If we define a kernel function as follows:

$$K(x_i, x_j) = \Phi(x_i).\Phi(x_j) \tag{A.35}$$

Then in the learning algorithm, it is enough to use only K without explicitly mentioning Φ. The above conditions indicate a linear separation but in higher dimensions. In [34], the author have shown that with this mapping, if the kernel function meets the conditions of the Mercer theorem, the kernel used is appropriate and the data will be inseparable in the main space in the written space. As a result, kernel replacement can create a nonlinear algorithm from the mentioned linear algorithm.

A.1.4 Support vector machines as multi-class separators

In the real world, many issues involve multi-class categories and are not limited to binary categories. Various solutions are offered for this category. Figure A.5 shows one of the simplest solutions that uses a binary

classification. Generally, a combination of binary methods is used for this classification. Another simple method is one against all, in which one category is assumed as positive data and the rest as negative [44, 40].

A.2 Conclusion

In this section, we reviewed support vector machines. SVM is used in a variety of applications from face recognition to speech recognition and more. Support vector method is a very efficient method in data classification, so it is one of the important tools in supervised machine learning. In addition to all its strengths, this method also has limitations. The biggest limitation of the support vector machines is the choice of Kernel. When the kernel is fixed, the SVM classification has only one parameter that can be changed by the user (the error penalty parameter). Choosing the best kernel for a particular issue is an important challenge. The second limitation is in speed and size in training and testing. Teaching very large databases is a numerically unsolvable problem. Solutions to such problems are provided, some of which were mentioned in the section. Discrete data poses another problem, although good results can be obtained with re-scaling. Another issue is the difficulty in designing a separator for multi-class SVM. This method is very good and high accuracy compared to other data classification methods such as Neural Network if the choices are made correctly (selection of quadratic solver, selection of Kernel, etc.) and in tasks such as pattern recognition and signature identification, high efficiency Has shown.

Appendix B

Fast Fourier Transform

For a signal f with N samples, the discrete Fourier transform (DFT) is written by:

$$X(k) = \Sigma_{n=0}^{N-1} X(n) \exp^{\frac{-2i\pi kn}{N}}, \qquad 0 \leq k < N \tag{B.1}$$

The Fast Fourier Transform (hereinafter denoted FFT) is simply a DFT calculated according to an algorithm making it possible to reduce the number of operations and, in particular, the number of multiplications to be performed. It should be noted, however, that reducing the number of arithmetic operations to be performed is not synonymous with reducing execution time. It all depends on the architecture of the processor that performs the processing. If we carry out the calculation directly without an efficient algorithm, we must carry out:

$$\begin{cases} \\ \text{N(N-1)} \qquad\qquad complexadditions \end{cases}$$

There are different FFT algorithms. The best known is surely that of Cooley-Tukey (also called temporal interlacing or "decimation in time") which reduces to

$$\frac{N}{2}\log_2(N) \qquad \text{the number of multiplications.}$$

There are two versions of the algorithm:

- FFT with time interleaving.
- FFT with frequency interleaving.

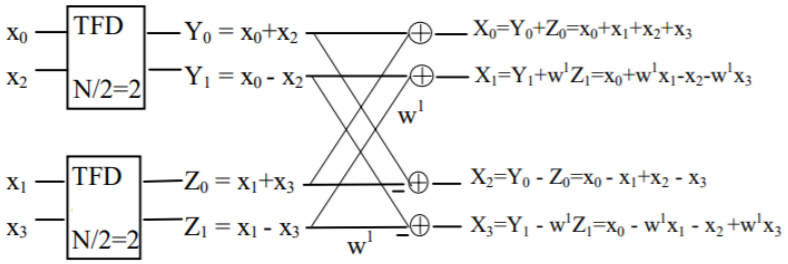
The algorithm requires that N be a power of 2. The principle of the algorithm consists in breaking down the computation of the DFT of order $N = 2^l$ into l successive steps.

B.0.1 FFT with time interleaving

Let us first illustrate the method by an example for $N = 4$. The data are denoted $x(n)$ and the TFD sequence $X(n)$. The notation w represents $\exp^{j2\pi/N}$. We can notice that $w^N = 1$ and $w^{N/2} = 1$. the TFD suite is written:

$$
\begin{aligned}
X(0) &= x(0)+x(1)+x(2)+x(3) = (x(0)+x(2))+(x(1)+x(3)) \\
X(1) &= x(0)+w^1x(1)+w^2x(2)+w^3x(3) = (x(0)-x(2))+w^1(x(1)-x(3)) \\
X(2) &= x(0)+w^2x(1)+w^4x(2)+w^6x(3) = (x(0)+x(2))-(x(1)+x(3)). \\
X(3) &= x(0)+w^3x(1)+w^6x(2)+w^9x(3) = (x(0)-x(2))-w^1(x(1)-x(3))
\end{aligned}
\tag{B.2}
$$

The data $(x(0),x(1),\ldots,x(N1))$ are grouped into 2 packets: a packet made up of data with even indices $(x(0),x(2),\ldots,x(N2))$ and a packet formed from data with odd indices $(x(1),x(3),\ldots,x(N1))$. Let for $N = 4$, a packet $(x(0),x(2))$ and a packet $(x(1),x(3))$. Then on each packet we perform a DFT of order $N/2$ and we combine the results of these 2 DFTs to obtain the one of order N. Which gives, again for $N = 4$:



To obtain the 4 values $X(k)$, it is therefore sufficient to calculate 2 DFT of order $N/2 = 2$ and to combine the results 2 to 2 using an addition and a multiplication at most, for each value $X(k)$. This step is called the "butterfly" stage, for obvious reasons linked to the shape of the calculation diagram. This result generalizes to any value of N multiple of 2. Indeed:

$$
\begin{aligned}
X(k) &= \Sigma_{n=0}^{N-1} x(n) e^{\frac{-2j\pi kn}{N}} \\
X(k) &= \Sigma_{i=0}^{N/2-1} x(2i) e^{\frac{-2j\pi 2ikn}{N}} + \Sigma_{i=0}^{N/2-1} x(2i+1) e^{\frac{-2j\pi 2(i+1)k}{N}} \\
X(k) &= \Sigma_{i=0}^{N/2-1} x(2i) e^{\frac{-j\pi ikn}{N/2}} + e^{-j2/N} \Sigma_{i=0}^{N/2-1} x(2i+1) e^{\frac{-2j\pi ik}{N/2}} \\
X(k) &= \Sigma_{i=0}^{N/2-1} y(i) e^{\frac{-j\pi ikn}{N/2}} + w^k \Sigma_{i=0}^{N/2-1} z(i) e^{\frac{-2j\pi (i+1)k}{N/2}}
\end{aligned}
\tag{B.3}
$$

We denote $y(i) = x(2i)$ and $z(i) = x(2i+1)$, for $i \in [0,(N/2-1)]$. Note that the 2 terms of the sum giving $X(k)$ can be deduced directly from the 2 DFTs of order $N/2$ of the sequences $y(i)$ and $z(i)$ of $N/2$ points. We denote these DFTs $Y(k)$ and $Z(k)$. Thus for $k \leq N/2-1$, the 2 terms of the sum are deduced from the terms of rank k of $Y(k)$ and $Z(k)$:

$$
X(k) = \Sigma_{i=0}^{N/2-1} y(i) e^{\frac{-j\pi ikn}{N/2}} + w^k \Sigma_{i=0}^{N/2-1} z(i) e^{\frac{-2j\pi (i+1)k}{N/2}} = Y(k) + w^k Z(k) \tag{B.4}
$$

For $k \in [N/2,(N-1)]$, we can write $k = k' + N/2$, with $k' \in [0,(N/2-1)]$. Moreover, since whatever i integer $e^{j2\pi i} = 1$, we can deduce $X(k)$ from the terms of rank $k-N/2$ of the 2 DFTs $Y(k)$ and $Z(k)$:

$$
\begin{aligned}
X(k) &= \Sigma_{i=0}^{N-1} y(i) e^{\frac{-2j\pi ik}{N/2}} + w^k \Sigma_{i=0}^{N/2-1} z(i) e^{\frac{-2j\pi ik}{N/2}} \\
X(k) &= \Sigma_{i=0}^{N-1} y(i) e^{\frac{-2j\pi i(k+N/2)}{N/2}} + w^k \Sigma_{i=0}^{N/2-1} z(i) e^{\frac{-2j\pi i(k+N/2)}{N/2}} \\
X(k) &= Y(k-N/2) + w^k Z(k-N/2)
\end{aligned}
\tag{B.5}
$$

In conclusion, for any N multiple of 2, we can calculate each term $X(k)$of the DFT of order N by combining, using at most 1 multiplication and 1 addition, 2 terms of the DFT of order $N/2$ of the 2 sequences $y(i)$ and $z(i)$ of length $N/2$, formed respectively of terms of even indexes and of odd index terms of the sequence $x(n)$. By noting $Y(k)$ and $Z(k)$ the TFD of order $N/2$ of these sequences, we can write:

$$\text{for} \quad k \in [0, \tfrac{N}{2} - 1] \qquad \begin{cases} X(k) = Y(k) + w^k Z(k) \\ X(k + N/2) = Y_k + w^{k+N/2} Z(k) = Y(k) w^k Z(k) \end{cases}$$

Each butterfly requires 1 multiplication and 2 addition or subtraction. Thus any DFT of order N multiple of 2, can be calculated using 2 DFT of order $N/2$ and a stage of $N/2$ butterflies. The computational complexity, for the DFT of order N is therefore equal to that of 2 DFT of order $N/2$ plus that of $N/2$ butterflies. If we suppose that the DFTs of order $N/2$ are calculated directly (without an efficient algorithm), we can say that: The calculation of an even DFT of order N, with this algorithm, requires:

$$\text{The calculation of 2 TFD of order } N/2\text{:} \quad \begin{cases} 2(\frac{N}{2})^2 = \frac{N^2}{2} \;\; complex \;\; multiplications \\ 2\frac{N}{2}(\frac{N}{2} - 1) = N(\frac{N}{2} - 1) \;\; complex \;\; additions \end{cases}$$

$$\text{The calculation of } N/2 \text{ butterflies:} \quad \begin{cases} \frac{N}{2} \;\; complex \;\; multiplications \\ N \;\; complex \;\; additions \end{cases}$$

Totaling

$$\begin{cases} \frac{N^2}{2} + \frac{N}{2} \;\; complex \;\; multiplications \\ \frac{N^2}{2} \;\; complex \;\; additions \end{cases}$$

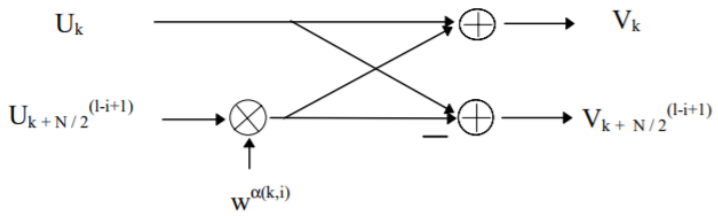
Instead of:

$$\begin{cases} N^2 \;\; complex \;\; multiplications \\ N(N-1) \;\; complex \;\; additions \end{cases}$$

For direct calculation. So for $N = 4$, we need 10 multiplications and 8 complex additions / subtractions instead of 16 multiplications and 12 complex additions / subtractions. If $N/2$ is a multiple of 2, we can repeat the method for calculating the 2 DFTs of order $N/2$. Each DFT of order $N/2$ is then calculated using 2 TFD of order $N/4$ and $N/4$ butterflies, which gives a total of 4 DFT of order $N/4$ plus 2 stages of $N/2$ butterflies. More generally if N is a power of 2, $N = 2^l$, we can repeat the method l times and calculate the DFT of order N using l stages of $N/2$ butterflies, with $l = log_2(N)$. The computational complexity of a DFT of order N then becomes that of l stages of $N/2$ butterflies, i.e:

$$\begin{cases} l\frac{N}{2} = log_2(N)^{\frac{N}{2}} \;\; complex \;\; multiplications \\ lN = log_2(N)N \;\; complex \;\; additions \end{cases}$$

This algorithm is the Cooley-Tukey time interleaving base 2 FFT algorithm. For the base 2 FFT algorithm with time interleaving, an elementary butterfly, at stage i (numbering from 1 to $l = log_2(N)$), has the following form:



In step i, the indices of the associated terms in a butterfly are separated from N_i, N_i being the size of the DFTs occurring in step i, i.e. $N_i = 2^{i1} = 2^l l/2^{(li+1)} = N/2^{(li+1)}$. The term $w^{\alpha(i,k)}$ is equal to:

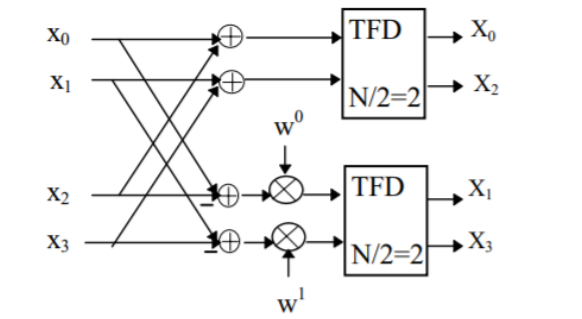
$$w^{\alpha(i,k)} = e^{-2j\pi\frac{k}{2N_i}} = e^{-j2\pi\frac{k}{N}\frac{N}{2^i}} = w^{k2^{l-i}} \tag{B.6}$$

B.1 FFT with frequency interleaving

This algorithm is symmetrical to the previous one. The $x(n)$ time data remains in the natural order, but the $X(k)$ results are disordered. The principle still consists in breaking down the calculation of the DFT of order $N = 2^l$ into l successive steps. But the grouping of data is done differently. Let us illustrate the method by an example for $N = 4$. The frequency data $(X(0), X(1), \ldots, X(N1))$ are grouped into 2 packets: a packet formed of index data even $(X(0), X(2), \ldots, X(N2))$ and a packet formed from data with odd indices $(X(1), X(3), \ldots, X((N1)))$. Let for $N = 4$, be a packet $(X(0), X(2))$ and a packet $(X(1), X(3))$. For $N = 4$, we can write:

$$\begin{aligned}
X(0) &= x(0) + x(1) + x(2) + x(3) = (x(0) + x(2)) + (x(1) + x(3)) \\
X(2) &= x(0) + w^2 x(1) + w^4 x(2) + w^6 x(3) = (x(0) + x(2)) - (x(1) + x(3)). \\
X(1) &= x(0) + w^1 x(1) + w^2 x(2) + w^3 x(3) = (x(0) - x(2)) + [w^1 (x(1) - x(3))] \\
X(3) &= x(0) + w^3 x(1) + w^6 x(2) + w^9 x(3) = (x(0) - x(2)) - [w^1 (x(1) - x(3))]
\end{aligned} \tag{B.7}$$

To obtain each packet of frequency results, we perform a DFT of order $N/2$ on data resulting from a butterfly step on the $x(n)$ data.



We therefore have a stage of 2 butterflies followed by a stage of 2 DFTs of order $N/2 = 2$. This result

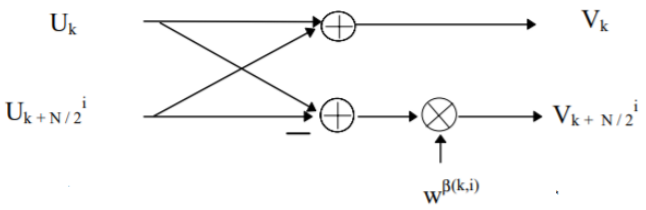
generalizes to any value of N multiple of 2. Indeed:

$$\begin{aligned} X(k) &= \Sigma_{n=0}^{N-1} x(n) e^{-j2\pi \frac{nk}{N}} \\ X(2i) &= \Sigma_{n=0}^{\frac{N}{2}-1} x(n) e^{-j2\pi \frac{n2i}{N}} + \Sigma_{n=\frac{N}{2}}^{N-1} x(n) e^{-j2\pi \frac{n2i}{N}} \\ X(2i) &= \Sigma_{n=0}^{\frac{N}{2}-1} x(n) e^{-j2\pi \frac{ni}{N/2}} + \Sigma_{m=O}^{N/2-1} x(m+\frac{N}{2}) e^{-j2\pi \frac{n2i}{N}} \\ X(2i) &= \Sigma_{n=0}^{\frac{N}{2}-1} (x(n)+x(n+\frac{N}{2})) e^{-j2\pi \frac{ni}{N/2}} \end{aligned} \tag{B.8}$$

Thus the $N/2$ terms $X(k)$ of even rank are equal to the terms of the DFT of order $N/2$ of the sequence of $N/2$ values $(x(n)+x(n+N/2))$, with n between 0 and $N/21$. Likewise for the terms $X(k)$ of odd rank:

$$\begin{aligned} X(k) &= \Sigma_n x(n) e^{-j2\pi \frac{nk}{N}} \\ X(2i+1) &= \Sigma_{n=0}^{\frac{N}{2}-1} x(n) e^{-j2\pi \frac{n(2i+)}{N}} + \Sigma_{n=\frac{N}{2}}^{N-1} x(n) e^{-j2\pi \frac{n(2i+1)}{N}} \\ X(2i+1) &= \Sigma_{n=0}^{\frac{N}{2}-1} x(n) e^{-j2\pi \frac{n}{N}} e^{-j\pi \frac{ni}{\frac{N}{2}}} - \Sigma_{m=O}^{N/2-1} x(m+\frac{N}{2}) e^{-j2\pi \frac{m}{N}} e^{-j\pi \frac{mi}{\frac{N}{2}}} \\ X(2i+1) &= \Sigma_{n=0}^{\frac{N}{2}-1} w^n (x(n)-x(n+\frac{N}{2})) e^{-j2\pi \frac{ni}{N/2}} \end{aligned} \tag{B.9}$$

The $N/2$ terms $X(k)$ of odd rank are equal to the terms of the DFT of order $N/2$ of the sequence of $N/2$ values $w^n(x(n)x(n+N/2))$, with n between 0 and $N/21$. In general, if N is a power of 2: $N = 2^l$, we can repeat the method l times and calculate the DFT of order N using l stages of $N/2$ butterflies, with $l = log_2(N)$. The computational complexity of an FFT with frequency interleaving is identical to that of FFT with time interleaving. For the base 2 FFT algorithm with frequency interleaving, an elementary butterfly, at stage i (by numbering from 1 to $l = log_2(N)$), has the following form:



In step i, the indices of associated terms in a butterfly are separated by N_i, N_i being the size of the DFTs occurring in step i, i.e $N_i = N/2^i$. And the term $w^{\beta(i,k)}$ is equal to:

$$w^{\beta(i,k)} = e^{-2j\pi \frac{k}{2N_i}} = e^{-j2\pi \frac{k2^i}{2N}} = w^{k2^{i1}} \tag{B.10}$$

Bibliography

[1] S. Agahian, F. Negin, and C. Köse. Improving bag-of-poses with semi-temporal pose descriptors for skeleton-based action recognition. The Visual Computer, 35(4):591–607, 2019.

[2] J. K. Aggarwal and M. S. Ryoo. Human activity analysis: A review. ACM Computing Surveys (CSUR), 43(3):1–43, 2011.

[3] M. Ahmad and S.-W. Lee. Variable silhouette energy image representations for recognizing human actions. Image and vision computing, 28(5):814–824, 2010.

[4] Z. Ahmad, K. Illanko, N. Khan, and D. Androutsos. Human action recognition using convolutional neural network and depth sensor data. In Proceedings of the 2019 International Conference on Information Technology and Computer Communications, pages 1–5, 2019.

[5] F. Ahmed and M. L. Gavrilova. Two-layer feature selection algorithm for recognizing human emotions from 3d motion analysis. In Computer Graphics International Conference, pages 53–67. Springer, 2019.

[6] I. Ajili, M. Mallem, and J.-Y. Didier. Human motions and emotions recognition inspired by lma qualities. The Visual Computer, 35(10):1411–1426, 2019.

[7] I. Ajili, Z. Ramezanpanah, M. Mallem, and J.-Y. Didier. Expressive motions recognition and analysis with learning and statistical methods. Multimedia Tools and Applications, 78(12):16575–16600, 2019.

[8] F. Angelini, Z. Fu, S. A. Velastin, J. A. Chambers, and S. M. Naqvi. 3d-hog embedding frameworks for single and multi-viewpoints action recognition based on human silhouettes. In 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 4219–4223. IEEE, 2018.

[9] A. Aristidou and Y. Chrysanthou. Feature extraction for human motion indexing of acted dance performances. In 2014 International Conference on Computer Graphics Theory and Applications (GRAPP), pages 1–11. IEEE, 2014.

[10] A. Aristidou, E. Stavrakis, and Y. Chrysanthou. Lma-based motion retrieval for folk dance cultural heritage. In Euro-Mediterranean Conference, pages 207–216. Springer, 2014.

[11] A. Aristidou, E. Stavrakis, and Y. Chrysanthou. Motion analysis for folk dance evaluation. In GCH, pages 55–64. Darmstadt, 2014.

[12] A. Aristidou, P. Charalambous, and Y. Chrysanthou. Emotion analysis and classification: understanding the performers' emotions using the lma entities. In Computer Graphics Forum, volume 34, pages 262–276. Wiley Online Library, 2015.

[13] A. Aristidou, E. Stavrakis, P. Charalambous, Y. Chrysanthou, and S. L. Himona. Folk dance evaluation using laban movement analysis. Journal on Computing and Cultural Heritage (JOCCH), 8(4):1–19, 2015.

[14] A. Aristidou, Q. Zeng, E. Stavrakis, K. Yin, D. Cohen-Or, Y. Chrysanthou, and B. Chen. Emotion control of unstructured dance movements. In Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation, pages 1–10, 2017.

[15] A. Aristidou, E. Stavrakis, M. Papaefthimiou, G. Papagiannakis, and Y. Chrysanthou. Style-based motion analysis for dance composition. The visual computer, 34(12):1725–1737, 2018.

[16] G. Assunção, P. Menezes, and F. Perdigão. Speaker awareness for speech emotion recognition. International Journal of Online and Biomedical Engineering (iJOE), 16(04):15–22, 2020.

[17] H. Aviezer, R. R. Hassin, J. Ryan, C. Grady, J. Susskind, A. Anderson, M. Moscovitch, and S. Bentin. Angry, disgusted, or afraid? studies on the malleability of emotion perception. Psychological science, 19 (7):724–732, 2008.

[18] A. M. Azab, H. Ahmadi, L. Mihaylova, and M. Arvaneh. Dynamic time warping-based transfer learning for improving common spatial patterns in brain–computer interface. Journal of Neural Engineering, 17 (1):016061, 2020.

[19] A. D. Bagdanov, A. Del Bimbo, and I. Masi. The florence 2d/3d hybrid face dataset. In Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding, pages 79–80, 2011.

[20] S. Balsis, L. D. Cooper, and T. F. Oltmanns. Are informant reports of personality more internally consistent than self reports of personality? Assessment, 22(4):399–404, 2015.

[21] E. I. Barakova and T. Lourens. Expressing and interpreting emotional movements in social games with robots. Personal and ubiquitous computing, 14(5):457–467, 2010.

[22] C. B. Barber, D. P. Dobkin, D. P. Dobkin, and H. Huhdanpaa. The quickhull algorithm for convex hulls. ACM Transactions on Mathematical Software (TOMS), 22(4):469–483, 1996.

[23] A. J. Barnes. Creating connection through dance/movement therapy among older adults with dementia: Development of a method. 2020.

[24] P. Barros, N. T. Maciel-Junior, B. J. Fernandes, B. L. Bezerra, and S. M. Fernandes. A dynamic gesture recognition and prediction system using the convexity approach. Computer Vision and Image Understanding, 155:139–149, 2017.

[25] I. Bartenieff, P. Hackney, B. T. Jones, J. Van Zile, and C. Wolz. The potential of movement analysis as a research tool: a preliminary analysis. Dance Research Journal, 16(1):3–26, 1984.

[26] E. Bauer and R. Kohavi. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine learning, 36(1-2):105–139, 1999.

[27] S. R. Bedico, E. M. L. Lope, E. J. L. Lope, E. B. Lunjas, A. P. D. Lustre, and R. E. Tolentino. Gesture recognition of basketball referee violation signal by applying dynamic time warping algorithm using a wearable device. In 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), pages 249–254. IEEE, 2020.

[28] R. Bellman and R. Kalaba. On adaptive control processes. IRE Transactions on Automatic Control, 4 (2):1–9, 1959.

[29] A. Ben Tanfous, H. Drira, and B. Ben Amor. Coding kendall's shape trajectories for 3d action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2840–2849, 2018.

[30] A. Bhavan, P. Chauhan, R. R. Shah, et al. Bagged support vector machines for emotion recognition from speech. Knowledge-Based Systems, 184:104886, 2019.

[31] C. M. Bishop and M. Tipping. Variational relevance vector machines. arXiv preprint arXiv:1301.3838, 2013.

[32] A. F. Bobick and J. W. Davis. Action recognition using temporal templates. In Motion-Based Recognition, pages 125–146. Springer, 1997.

[33] A. F. Bobick and J. W. Davis. The recognition of human movement using temporal templates. IEEE Transactions on pattern analysis and machine intelligence, 23(3):257–267, 2001.

[34] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual workshop on Computational learning theory, pages 144–152, 1992.

[35] L. Breiman. Random forests. Machine learning, 45(1):5–32, 2001.

[36] L. Breiman, J. Friedman, R. Olshen, and C. Stone. Classification and regression trees, wadsworth international group, belmont, california, usa, 1984; bp roe et al., boosted decision trees as an alternative to artificial neural networks for particle identification. Nucl. Instrum. Meth. A, 543(577):10–1016, 2005.

[37] P. Breuer, C. Eckes, and S. Müller. Hand gesture recognition with a novel ir time-of-flight range camera–a pilot study. In International Conference on Computer Vision/Computer Graphics Collaboration Techniques and Applications, pages 247–260. Springer, 2007.

[38] A. Bronstein, M. Bronstein, U. Castellani, A. Dubrovina, L. Guibas, R. Horaud, R. Kimmel, D. Knossow, E. Von Lavante, D. Mateus, et al. Shrec 2010: robust correspondence benchmark. In Eurographics Workshop on 3D Object Retrieval, volume 10, pages 087–091, 2010.

[39] M. Broughton and C. Stevens. Music, movement and marimba: An investigation of the role of movement and gesture in communicating musical expression to an audience. Psychology of Music, 37(2):137–153, 2009.

[40] C. J. Burges. A tutorial on support vector machines for pattern recognition. Data mining and knowledge discovery, 2(2):121–167, 1998.

[41] J. Butepage, H. Kjellstrom, and D. Kragic. Predicting the what and how-a probabilistic semi-supervised approach to multi-task human activity modeling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 0–0, 2019.

[42] T. Bylander. Estimating generalization error on two-class datasets using out-of-bag estimates. Machine learning, 48(1-3):287–297, 2002.

[43] C. Cadoz. Le geste canal de communication homme/machine: la communication" instrumentale". Technique et science informatiques, 13(1):31–61, 1994.

[44] C. Campbell. Kernel methods: a survey of current techniques. Neurocomputing, 48(1-4):63–84, 2002.

[45] A. Camurri, S. Hashimoto, M. Ricchetti, A. Ricci, K. Suzuki, R. Trocca, and G. Volpe. Eyesweb: Toward gesture and affect recognition in interactive dance and music systems. Computer Music Journal, 24(1): 57–69, 2000.

[46] A. Camurri, B. Mazzarino, M. Ricchetti, R. Timmers, and G. Volpe. Multimodal analysis of expressive gesture in music and dance performances. In International gesture workshop, pages 20–39. Springer, 2003.

[47] A. Camurri, P. Coletta, A. Massari, B. Mazzarino, M. Peri, M. Ricchetti, A. Ricci, and G. Volpe. Toward real-time multimodal processing: Eyesweb 4.0. In Proc. AISB. Citeseer, 2004.

[48] L. Cao, Y. Tian, Z. Liu, B. Yao, Z. Zhang, and T. S. Huang. Action detection using multiple spatial-temporal interest point features. In 2010 IEEE International Conference on Multimedia and Expo, pages 340–345. IEEE, 2010.

[49] W. Cao, Y. Lu, and Z. He. Geometric algebra representation and ensemble action classification method for 3d skeleton orientation data. IEEE Access, 7:132049–132056, 2019.

[50] W. Cao, J. Zhong, G. Cao, and Z. He. Physiological function assessment based on kinect v2. IEEE Access, 7:105638–105651, 2019.

[51] B. Chakraborty, M. B. Holte, T. B. Moeslund, and J. Gonzàlez. Selective spatio-temporal interest points. Computer Vision and Image Understanding, 116(3):396–410, 2012.

[52] C.-C. Chang. " libsvm: a library for support vector machines," acm transactions on intelligent systems and technology, 2: 27: 1–27: 27, 2011. http://www. csie. ntu. edu. tw/˜ cjlin/libsvm, 2, 2011.

[53] D. Chi, M. Costa, L. Zhao, and N. Badler. The emote model for effort and shape. In Proceedings of the 27th annual conference on Computer graphics and interactive techniques, pages 173–182, 2000.

[54] F. Chin-Shyurng, S.-E. Lee, and M.-L. Wu. Real-time musical conducting gesture recognition based on a dynamic time warping classifier using a single-depth camera. Applied Sciences, 9(3):528, 2019.

[55] S. Chu, E. Keogh, D. Hart, and M. Pazzani. Iterative deepening dynamic time warping for time series. In Proceedings of the 2002 SIAM International Conference on Data Mining, pages 195–212. SIAM, 2002.

[56] M. ComaniciuD. Arobust approach toward feature space analysis [j]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(5):313–329, 2002.

[57] B. S. Connelly and D. S. Ones. An other perspective on personality: Meta-analytic integration of observers' accuracy and predictive validity. Psychological bulletin, 136(6):1092, 2010.

[58] A. S. Cowen, P. Laukka, H. A. Elfenbein, R. Liu, and D. Keltner. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. Nature human behaviour, 3(4):369–382, 2019.

[59] H. Cui, C. Maguire, and A. LaViers. Laban-inspired task-constrained variable motion generation on expressive aerial robots. Robotics, 8(2):24, 2019.

[60] Y. Dai and J. Zhao. Fault diagnosis of batch chemical processes using a dynamic time warping (dtw)-based artificial immune system. Industrial & Engineering Chemistry Research, 50(8):4534–4544, 2011.

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

[62] R. J. Davidson, P. Ekman, C. D. Saron, J. A. Senulis, and W. V. Friesen. Approach-withdrawal and cerebral asymmetry: emotional expression and brain physiology: I. Journal of personality and social psychology, 58(2):330, 1990.

[63] B. De Gelder. Towards the neurobiology of emotional body language. Nature Reviews Neuroscience, 7 (3):242–249, 2006.

[64] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo. 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold. IEEE transactions on cybernetics, 45(7):1340–1352, 2014.

[65] T. G. Dietterich. Ensemble methods in machine learning. In International workshop on multiple classifier systems, pages 1–15. Springer, 2000.

[66] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. VS-PETS Beijing, China, 2005.

[67] D. Droeschel, J. Stückler, and S. Behnke. Learning to interpret pointing gestures with a time-of-flight camera. In Proceedings of the 6th international conference on Human-robot interaction, pages 481–488. ACM, 2011.

[68] G.-B. Duchenne and G.-B. D. de Boulogne. The mechanism of human facial expression. Cambridge university press, 1990.

[69] F. Durupinar, M. Kapadia, S. Deutsch, M. Neff, and N. I. Badler. Perform: Perceptual approach for adding ocean personality to human motion using laban movement analysis. ACM Transactions on Graphics (TOG), 36(1):1–16, 2016.

[70] P. Ekman, D. Matsumoto, and W. V. Friesen. Facial expression in affective disorders. What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS), 2:331–342, 1997.

[71] A. Entezami and H. Shariatmadar. Structural health monitoring by a new hybrid feature extraction and dynamic time warping methods under ambient vibration and non-stationary signals. Measurement, 134: 548–568, 2019.

[72] C.-H. Fang, J.-C. Chen, C.-C. Tseng, and J.-J. J. Lien. Human action recognition using spatio-temporal classification. In Asian Conference on Computer Vision, pages 98–109. Springer, 2009.

[73] Š. Fojtu and D. Pruša. Demo applications for humanoid robot asterix. 2010.

[74] A. Foroud and I. Q. Whishaw. Changes in the kinematic structure and non-kinematic features of movements during skilled reaching after stroke: A laban movement analysis in two case studies. Journal of neuroscience methods, 158(1):137–149, 2006.

[75] N. Fourati and C. Pelachaud. Emilya: Emotional body expression in daily actions database. In LREC, pages 3486–3493, 2014.

[76] N. Fourati and C. Pelachaud. Perception of emotions and body movement in the emilya database. IEEE Transactions on Affective Computing, 9(1):90–101, 2016.

[77] N. Fourati, C. Pelachaud, and P. Darmon. Contribution of temporal and multi-level body cues to emotion classification. In 2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII), pages 116–122. IEEE, 2019.

[78] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1):119–139, 1997.

[79] N. H. Frijda. What might emotions be? comments on the comments. Social Science Information, 46(3): 433–443, 2007.

[80] G. Fumera and F. Roli. A theoretical and experimental analysis of linear combiners for multiple classifier systems. IEEE transactions on pattern analysis and machine intelligence, 27(6):942–956, 2005.

[81] M. Garber-Barron and M. Si. Using body movement and posture for emotion detection in non-acted scenarios. In 2012 IEEE International Conference on Fuzzy Systems, pages 1–8. IEEE, 2012.

[82] G. Garcia-Hernando and T.-K. Kim. Transition forests: Learning discriminative temporal transitions for action recognition and detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 432–440, 2017.

[83] D. M. Gavrila, L. S. Davis, et al. Towards 3-d model-based tracking and recognition of human movement: a multi-view approach. In International workshop on automatic face-and gesture-recognition, volume 3, pages 272–277. Citeseer, 1995.

[84] F. A. Gers, J. Schmidhuber, and F. Cummins. Learning to forget: Continual prediction with lstm. 1999.

[85] E. Ghorbel. Fast and accurate human action recognition using RGB-D cameras. PhD thesis, 2017.

[86] D. Glowinski, N. Dael, A. Camurri, G. Volpe, M. Mortillaro, and K. Scherer. Toward a minimal representation of affective gestures. IEEE Transactions on Affective Computing, 2(2):106–118, 2011.

[87] E. Groff. Laban movement analysis: Charting the ineffable domain of human movement. Journal of Physical Education, Recreation & Dance, 66(2):27–30, 1995.

[88] A. H. Guest. Labanotation, Or, Kinetography Laban: The System of Analyzing and Recording Movement. Number 27. Dance Books, 1996.

[89] K. Guo, P. Ishwar, and J. Konrad. Action recognition in video by covariance matching of silhouette tunnels. In 2009 XXII Brazilian Symposium on Computer Graphics and Image Processing, pages 299–306. IEEE, 2009.

[90] K. Hachimura, K. Takashina, and M. Yoshimura. Analysis and evaluation of dancing movement based on lma. In ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005., pages 294–299. IEEE, 2005.

[91] S. Halovic and C. Kroos. Not all is noticed: Kinematic cues of emotion-specific gait. Human movement science, 57:478–488, 2018.

[92] K. Hara, H. Kataoka, and Y. Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 3154–3160, 2017.

[93] C. G. Harris, M. Stephens, et al. A combined corner and edge detector. In Alvey vision conference, volume 15, pages 10–5244. Citeseer, 1988.

[94] B. Hartmann, M. Mancini, and C. Pelachaud. Implementing expressive gesture synthesis for embodied conversational agents. In International Gesture Workshop, pages 188–199. Springer, 2005.

[95] S. Haykin and N. Network. A comprehensive foundation. Neural networks, 2(2004):41, 2004.

[96] S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997.

[97] S. Hochreiter and J. Schmidhuber. Lstm can solve hard long time lag problems. In Advances in neural information processing systems, pages 473–479, 1997.

[98] M. B. Holte, T. B. Moeslund, and P. Fihl. View invariant gesture recognition using the csem swissranger sr-2 camera. IJISTA, 5(3/4):295–303, 2008.

[99] M. Hoque and M. Hannan. Performance evaluation of laser guided leveler. International Journal of Agricultural Research, Innovation and Technology, 4(2):82–86, 2014.

[100] W. Hou, Q. Pan, Q. Peng, and M. He. A new method to analyze protein sequence similarity using dynamic time warping. Genomics, 109(2):123–130, 2017.

[101] E. Hsu, K. Pulli, and J. Popović. Style translation for human motion. In ACM SIGGRAPH 2005 Papers, pages 1082–1089. 2005.

[102] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang. Jointly learning heterogeneous features for rgb-d activity recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 5344–5352, 2015.

[103] J.-F. Hu, W.-S. Zheng, J. Pan, J. Lai, and J. Zhang. Deep bilinear learning for rgb-d action recognition. In Proceedings of the European Conference on Computer Vision (ECCV), pages 335–351, 2018.

[104] L. Huang, Y. Huang, W. Ouyang, L. Wang, et al. Part-level graph convolutional network for skeleton-based action recognition. 2020.

[105] C. J. Huberty. Applied discriminant analysis. Number 519.535 HUB. CIMMYT. 1994.

[106] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In Twenty-Third International Joint Conference on Artificial Intelligence, 2013.

[107] R. R. Igorevich, P. Park, J. Choi, and D. Min. Two hand gesture recognition using stereo camera. International Journal of Computer and Electrical Engineering, 5(1):69–72, 2013.

[108] S. M. Islam, A. Rahman, N. Prasad, O. Boric-Lubecke, and V. M. Lubecke. Identity authentication system using a support vector machine (svm) on radar respiration measurements. In 2019 93rd ARFTG Microwave Measurement Conference (ARFTG), pages 1–5. IEEE, 2019.

[109] C. E. Izard. Basic emotions, natural kinds, emotion schemas, and a new paradigm. Perspectives on psychological science, 2(3):260–280, 2007.

[110] S. Jazayeri, A. Saghafi, S. Esmaeili, and C. P. Tsokos. Automatic object detection using dynamic time warping on ground penetrating radar signals. Expert Systems with Applications, 122:102–107, 2019.

[111] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black. Towards understanding action recognition. In Proceedings of the IEEE international conference on computer vision, pages 3192–3199, 2013.

[112] X. Jiang, F. Zhong, Q. Peng, and X. Qin. Online robust action recognition based on a hierarchical model. The Visual Computer, 30(9):1021–1033, 2014.

[113] G. Johansson. Visual perception of biological motion and a model for its analysis. Perception & psychophysics, 14(2):201–211, 1973.

[114] I. N. Junejo, K. N. Junejo, and Z. Al Aghbari. Silhouette-based human action recognition using sax-shapes. The Visual Computer, 30(3):259–269, 2014.

[115] A. Kacem. Novel Geometric Tools for Human Behavior Understanding. PhD thesis, 2018.

[116] M. Kapadia, I.-k. Chiang, T. Thomas, N. I. Badler, and J. T. Kider Jr. Efficient motion retrieval in large motion databases. In Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games, pages 19–28, 2013.

[117] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017.

[118] Q. Ke, M. Bennamoun, S. An, F. Sohel, and F. Boussaid. A new representation of skeleton sequences for 3d action recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3288–3297, 2017.

[119] D. G. Kendall. Shape manifolds, procrustean metrics, and complex projective spaces. Bulletin of the London mathematical society, 16(2):81–121, 1984.

[120] A. Kendon. Gesture: Visible action as utterance. Cambridge University Press, 2004.

[121] E. Keogh, B. Celly, C. A. Ratanamahatana, and V. B. Zordan. A novel technique for indexing video surveillance data. In First ACM SIGMM international workshop on Video surveillance, pages 98–106, 2003.

[122] E. J. Keogh and M. J. Pazzani. Scaling up dynamic time warping for datamining applications. In Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, pages 285–289, 2000.

[123] J. Kim, J.-H. Seo, and D.-S. Kwon. Application of effort parameter to robot gesture motion. In 2012 9th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), pages 80–82. IEEE, 2012.

[124] M.-G. Kim, E. Barakova, and T. Lourens. Rapid prototyping framework for robot-assisted training of autistic children. In The 23rd IEEE International Symposium on Robot and Human Interactive Communication, pages 353–358. IEEE, 2014.

[125] W. Kim, J. Lee, M. Kim, D. Oh, and C. Kim. Human action recognition using ordinal measure of accumulated motion. EURASIP journal on Advances in Signal Processing, 2010(1):1–11, 2010.

[126] W. H. Kim, J. W. Park, W. H. Lee, M. J. Chung, and H. S. Lee. Lma based emotional motion representation using rgb-d camera. In 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pages 163–164. IEEE, 2013.

[127] D. Kisku, M. Tistarelli, and J. Sing. Computer vision and pattern recognition workshops. Miami, Florida, USA, page 60, 2009.

[128] A. Klaser, M. Marszałek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. 2008.

[129] H. Knight and R. Simmons. Expressive motion with x, y and theta: Laban effort features for mobile robots. In The 23rd IEEE International Symposium on Robot and Human Interactive Communication, pages 267–273. IEEE, 2014.

[130] H. Knight and R. Simmons. Laban head-motions convey robot state: A call for robot body language. In 2016 IEEE international conference on robotics and automation (ICRA), pages 2881–2888. IEEE, 2016.

[131] H. Knight, R. Thielstrom, and R. Simmons. Expressive path shape (swagger): Simple features that illustrate a robot's attitude toward its goal in real time. In 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 1475–1482. IEEE, 2016.

[132] Y. Kong, B. Satarboroujeni, and Y. Fu. Learning hierarchical 3d kernel descriptors for rgb-d action recognition. Computer Vision and Image Understanding, 144:14–23, 2016.

[133] P. Koniusz, A. Cherian, and F. Porikli. Tensor representations via kernel linearization for action recognition from 3d skeletons. In European Conference on Computer Vision, pages 37–53. Springer, 2016.

[134] R. Laban and L. Ullmann. The mastery of movement. 1971.

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

[136] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2008.

[137] C. Larboulette and S. Gibet. A review of computable expressive descriptors of human motion. In Proceedings of the 2nd International Workshop on Movement and Computing, pages 21–28, 2015.

[138] I. Lee, D. Kim, S. Kang, and S. Lee. Ensemble deep learning for skeleton-based action recognition using temporal sliding lstm networks. In Proceedings of the IEEE international conference on computer vision, pages 1012–1020, 2017.

[139] Y.-J. Lee and S.-Y. Huang. Reduced support vector machines: A statistical theory. IEEE Transactions on neural networks, 18(1):1–13, 2007.

[140] M. Leman, A. CAMURRI, and G. DE POLI. Megase: a multisensory expressive gesture applications system environment for artistic performances. In Conference on Communication of Art, Science and Technology, 2001.

[141] H. Li, J. Liu, Z. Yang, R. W. Liu, K. Wu, and Y. Wan. Adaptively constrained dynamic time warping for time series classification and clustering. Information Sciences, 2020.

[142] M. Li, H. Leung, and H. P. Shum. Human action recognition via skeletal and depth based feature fusion. In Proceedings of the 9th International Conference on Motion in Games, pages 123–132, 2016.

[143] M. Li, L. Yan, and Q. Wang. Group sparse regression-based learning model for real-time depth-based human action prediction. Mathematical Problems in Engineering, 2018, 2018.

[144] M. Li, S. Chen, X. Chen, Y. Zhang, Y. Wang, and Q. Tian. Actional-structural graph convolutional networks for skeleton-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3595–3603, 2019.

[145] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3d points. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pages 9–14. IEEE, 2010.

[146] X. Li, Y. Zhang, and D. Liao. Mining key skeleton poses with latent svm for action recognition. Applied Computational Intelligence and Soft Computing, 2017, 2017.

[147] Y. Li, Y. Wang, Y. Jiang, and L. Zhang. Action recognition using convolutional neural networks with joint supervision. In 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pages 2015–2020. IEEE, 2019.

[148] Y. Li, R. Xia, X. Liu, and Q. Huang. Learning shape-motion representations from geometric algebra spatio-temporal model for skeleton-based action recognition. In 2019 IEEE International Conference on Multimedia and Expo (ICME), pages 1066–1071. IEEE, 2019.

[149] Z. Li, J. Guo, H. Li, T. Wu, S. Mao, and F. Nie. Speed up similarity search of time series under dynamic time warping. IEEE Access, 7:163644–163653, 2019.

[150] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzel. Learning to diagnose with lstm recurrent neural networks. arXiv preprint arXiv:1511.03677, 2015.

[151] J. Liu, A. Shahroudy, M. L. Perez, G. Wang, L.-Y. Duan, and A. K. Chichung. Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. IEEE transactions on pattern analysis and machine intelligence, 2019.

[152] M. Liu and J. Yuan. Recognizing human actions as the evolution of pose estimation maps. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1159–1168, 2018.

[153] M. Liu, H. Liu, and C. Chen. Enhanced skeleton visualization for view invariant human action recognition. Pattern Recognition, 68:346–362, 2017.

[154] E. Lorini and F. Schwarzentruber. A logic for reasoning about counterfactual emotions. Artificial Intelligence, 175(3-4):814–847, 2011.

[155] T. Lourens, R. Van Berkel, and E. Barakova. Communicating emotions and mental states to robots in a real time parallel framework using laban movement analysis. Robotics and Autonomous Systems, 58(12): 1256–1265, 2010.

[156] Y. Luo, J. Ye, R. B. Adams, J. Li, M. G. Newman, and J. Z. Wang. Arbee: Towards automated recognition of bodily expression of emotion in the wild. International Journal of Computer Vision, 128(1):1–25, 2020.

[157] V. Maletic and S. Body. Expression: The development of rudolf laban's movement and dance concepts. New York: Mouton de Gruyter, 1987.

[158] T. Marill and D. Green. On the effectiveness of receptors in recognition systems. IEEE transactions on Information Theory, 9(1):11–17, 1963.

[159] M. Masuda and S. Kato. Motion rendering system for emotion expression of human form robots based on laban movement analysis. In 19Th international symposium in robot and human interactive communication, pages 324–329. IEEE, 2010.

[160] M. Masuda, S. Kato, and H. Itoh. A laban-based approach to emotional motion rendering for human-robot interaction. In International Conference on Entertainment Computing, pages 372–380. Springer, 2010.

[161] MATLAB. 9.7.0.1190202 (R2019b). The MathWorks Inc., Natick, Massachusetts, 2018.

[162] G. McLaine and S. Lexow. Labanotation and lma: The symbiotic relationship in dance reconstruction. Journal of Dance Education, 20(1):44–47, 2020.

[163] D. McNeill. Hand and mind: What gestures reveal about thought. University of Chicago press, 1992.

[164] A. Mehrabian and J. T. Friar. Encoding of attitude by a seated communicator via posture and position cues. Journal of Consulting and Clinical Psychology, 33(3):330, 1969.

[165] F. Michaud, E. Robichaud, and J. Audet. Using motives and artificial emotions for prolonged activity of a group of autonomous robots. In Proceedings of the AAAI Fall Symposium on Emotions. Cape Code Massachussetts, 2001.

[166] M. Y. Y. MOHAMMED and M. CELIK. Developing fast techniques for periodicity analysis of time series. In 2019 3rd International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), pages 1–5. IEEE, 2019.

[167] M. Morel, C. Achard, R. Kulpa, and S. Dubuisson. Time-series averaging using constrained dynamic time warping with tolerance. Pattern Recognition, 74:77–89, 2018.

[168] N. Moshtagh et al. Minimum volume enclosing ellipsoid. Convex optimization, 111(January):1–9, 2005.

[169] M. Müller. Dtw-based motion comparison and retrieval. Information Retrieval for Music and Motion, pages 211–226, 2007.

[170] M. Müller, H. Mattes, and F. Kurth. An efficient multiscale approach to audio synchronization. In ISMIR, volume 546, pages 192–197. Citeseer, 2006.

[171] R. M. Murray, Z. Li, S. S. Sastry, and S. S. Sastry. A mathematical introduction to robotic manipulation. CRC press, 1994.

[172] C. Myers, L. Rabiner, and A. Rosenberg. Performance tradeoffs in dynamic time warping algorithms for isolated word recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(6):623–635, 1980.

[173] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. Applied linear statistical models. 1996.

[174] K. Nishimura, N. Kubota, and J. Woo. Design support system for emotional expression of robot partners using interactive evolutionary computation. In 2012 IEEE International Conference on Fuzzy Systems, pages 1–7. IEEE, 2012.

[175] J. C. Nunez, R. Cabido, J. J. Pantrigo, A. S. Montemayor, and J. F. Velez. Convolutional neural networks and long short-term memory for skeleton-based human activity and hand gesture recognition. Pattern Recognition, 76:80–94, 2018.

[176] J. C. Nunnally. Psychometric theory—25 years ago and now. Educational Researcher, 4(10):7–21, 1975.

[177] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Berkeley mhad: A comprehensive multimodal human action database. In 2013 IEEE Workshop on Applications of Computer Vision (WACV), pages 53–60. IEEE, 2013.

[178] E. J. Okafor. Music Retrieval System Using Dynamic Time Warping. PhD thesis, North Carolina Agricultural and Technical State University, 2019.

[179] O. Oreifej and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 716–723, 2013.

[180] K. Ouivirach and M. N. Dailey. Clustering human behaviors with dynamic time warping and hidden markov models for a video surveillance system. In ECTI-CON2010: The 2010 ECTI International Confernce on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, pages 884–888. IEEE, 2010.

[181] K. O'regan. Emotion and e-learning. Journal of Asynchronous learning networks, 7(3):78–92, 2003.

[182] K. Papadopoulos, E. Ghorbel, D. Aouada, and B. Ottersten. Vertex feature encoding and hierarchical temporal modeling in a spatial-temporal graph convolutional network for action recognition. arXiv preprint arXiv:1912.09745, 2019.

[183] A. Parziale, M. Diaz, M. A. Ferrer, and A. Marcelli. Sm-dtw: Stability modulated dynamic time warping for signature verification. Pattern Recognition Letters, 121:113–122, 2019.

[184] F. Pedersoli, S. Benini, N. Adami, and R. Leonardi. Xkin: an open source framework for hand pose and gesture recognition using kinect. The Visual Computer, 30(10):1107–1122, 2014.

[185] L. Pei, M. Ye, X. Zhao, Y. Dou, and J. Bao. Action recognition by learning temporal slowness invariant features. The Visual Computer, 32(11):1395–1404, 2016.

[186] C. Peter and R. Beale. Affect and emotion in human-computer interaction: From theory to applications, volume 4868. Springer Science & Business Media, 2008.

[187] H.-H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin. Exploiting deep residual networks for human action recognition from skeletal data. Computer Vision and Image Understanding, 170:51–66, 2018.

[188] H. H. Pham, H. Salmane, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin. A deep learning approach for real-time 3d human action recognition from skeletal data. In International Conference on Image Analysis and Recognition, pages 18–32. Springer, 2019.

[189] R. W. Picard. Affective computing: challenges. International Journal of Human-Computer Studies, 59 (1-2):55–64, 2003.

[190] A. Pikrakis, S. Theodoridis, and D. Kamarotos. Recognition of isolated musical patterns using context dependent dynamic time warping. IEEE Transactions on Speech and Audio Processing, 11(3):175–183, 2003.

[191] J. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.

[192] G. Plouffe and A.-M. Cretu. Static and dynamic hand gesture recognition in depth data using dynamic time warping. IEEE transactions on instrumentation and measurement, 65(2):305–316, 2015.

[193] A. E. Poropat. Other-rated personality and academic performance: Evidence and implications. Learning and Individual Differences, 34:24–32, 2014.

[194] S. Pramanick. Using dynamic time warping to improve the classical music production workflow. PhD thesis, Massachusetts Institute of Technology, 2019.

[195] C. QIAN, J. SHAO, T. XIA, and H. LIU. Method for sign language recognition based on kinect. Transducer and Microsystem Technologies, (6):9, 2019.

[196] J. R. Quinlan et al. Learning with continuous classes. In 5th Australian joint conference on artificial intelligence, volume 92, pages 343–348. World Scientific, 1992.

[197] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian. Histogram of oriented principal components for cross-view action recognition. IEEE transactions on pattern analysis and machine intelligence, 38(12): 2430–2443, 2016.

[198] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh. Searching and mining trillions of time series subsequences under dynamic time warping. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 262–270, 2012.

[199] Z. Ramezanpanah, M. Mallem, and F. Davesne. Human action recognition using laban movement analysis and dynamic time warping. In 24th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2020), 2020.

[200] C. E. Rasmussen and H. Nickisch. Gaussian processes for machine learning (gpml) toolbox. Journal of machine learning research, 11(Nov):3011–3015, 2010.

[201] M. A. Razzaq, J. Bang, S. S. Kang, and S. Lee. Unskem: Unobtrusive skeletal-based emotion recognition for user experience. In 2020 International Conference on Information Networking (ICOIN), pages 92–96. IEEE, 2020.

[202] M. Rhif, H. Wannous, and I. R. Farah. Action recognition from 3d skeleton sequences using deep networks on lie group features. In 2018 24th International Conference on Pattern Recognition (ICPR), pages 3427–3432. IEEE, 2018.

[203] A. Robotics. Project romeo, 2010.

[204] D. Roetenberg, H. Luinge, and P. Slycke. Xsens mvn: Full 6dof human motion tracking using miniature inertial sensors. Xsens Motion Technologies BV, Tech. Rep, 1, 2009.

[205] L. Rokach. Pattern classification using ensemble methods, volume 75. World Scientific, 2010.

[206] J. A. Russell. A circumplex model of affect. Journal of personality and social psychology, 39(6):1161, 1980.

[207] G. Saggio, P. Cavallo, M. Ricci, V. Errico, J. Zea, and M. E. Benalcázar. Sign language recognition using wearable electronics: Implementing k-nearest neighbors with dynamic time warping and convolutional neural network algorithms. Sensors, 20(14):3879, 2020.

[208] S. P. Sahoo, R. Silambarasi, and S. Ari. Fusion of histogram based features for human action recognition. In 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), pages 1012–1016. IEEE, 2019.

[209] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing, 26(1):43–49, 1978.

[210] A. Samadani, R. Gorbet, and D. Kulic. Affective movement generation using laban effort and shape and hidden markov models. arXiv preprint arXiv:2006.06071, 2020.

[211] A.-A. Samadani, S. Burton, R. Gorbet, and D. Kulic. Laban effort and shape analysis of affective hand and arm movements. In 2013 Humaine Association conference on affective computing and intelligent interaction, pages 343–348. IEEE, 2013.

[212] R. Santhoshkumar and M. K. Geetha. Human emotion recognition using body expressive feature. In Microservices in Big Data Analytics, pages 141–149. Springer, 2020.

[213] M. Schepers, M. Giuberti, G. Bellusci, et al. Xsens mvn: Consistent tracking of human motion using inertial sensing. Xsens Technologies, pages 1–8, 2018.

[214] K. R. Scherer. Vocal communication of emotion: A review of research paradigms. Speech communication, 40(1-2):227–256, 2003.

[215] K. R. Scherer. What are emotions? and how can they be measured? Social science information, 44(4): 695–729, 2005.

[216] P. Schneider, R. Memmesheimer, I. Kramer, and D. Paulus. Gesture recognition in rgb videos using human body keypoints and dynamic time warping. In Robot World Cup, pages 281–293. Springer, 2019.

[217] P. Schroeder. Distinctive image features from scale-invariant keypoints. Retrieved at, page 16, 2008.

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

[219] B. Schuller, S. Reiter, and G. Rigoll. Evolutionary feature generation in speech emotion recognition. In 2006 IEEE International Conference on Multimedia and Expo, pages 5–8. IEEE, 2006.

[220] L. Seidenari, V. Varano, S. Berretti, A. Bimbo, and P. 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.

[221] S. Senecal, L. Cuel, A. Aristidou, and N. Magnenat-Thalmann. Continuous body emotion recognition system during theater performances. Computer Animation and Virtual Worlds, 27(3-4):311–320, 2016.

[222] S. Senecal, N. A. Nijdam, A. Aristidou, and N. Magnenat-Thalmann. Salsa dance learning evaluation and motion analysis in gamified virtual reality environment. Multimedia Tools and Applications, pages 1–23, 2020.

[223] V. Sethu, E. Ambikairajah, and J. Epps. Speaker normalisation for speech-based emotion detection. In 2007 15th international conference on digital signal processing, pages 611–614. IEEE, 2007.

[224] A. Shahroudy, J. Liu, T.-T. Ng, and G. 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.

[225] L. Shao and X. Chen. Histogram of body poses and spectral regression discriminant analysis for human action categorization. In BMVC, pages 1–11, 2010.

[226] M. Sharma, D. Hildebrandt, G. Newman, J. E. Young, and R. Eskicioglu. Communicating affect via flight path exploring use of the laban effort system for designing affective locomotion paths. In 2013 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI), pages 293–300. IEEE, 2013.

[227] R. P. Sharma and G. K. Verma. Human computer interaction using hand gesture. Procedia Computer Science, 54:721–727, 2015.

[228] L. Shi, Y. Zhang, J. Cheng, and H. Lu. Skeleton-based action recognition with directed graph neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7912–7921, 2019.

[229] C. Si, W. Chen, W. Wang, L. Wang, and T. Tan. An attention enhanced graph convolutional lstm network for skeleton-based action recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1227–1236, 2019.

[230] R. Silambarasi, S. P. Sahoo, and S. Ari. 3d spatial-temporal view based motion tracing in human action recognition. In 2017 International Conference on Communication and Signal Processing (ICCSP), pages 1833–1837. IEEE, 2017.

[231] R. Slama, H. Wannous, M. Daoudi, and A. Srivastava. Accurate 3d action recognition using learning on the grassmann manifold. Pattern Recognition, 48(2):556–567, 2015.

[232] T. Sobol-Shikler and P. Robinson. Classification of complex information: Inference of co-occurring affective states from their expressions in speech. IEEE Transactions on Pattern Analysis and Machine Intelligence, 32(7):1284–1297, 2009.

[233] L. Song and M. Takatsuka. Real-time 3d finger pointing for an augmented desk. In Proceedings of the Sixth Australasian conference on User interface-Volume 40, pages 99–108. Australian Computer Society, Inc., 2005.

[234] Y. Song, J. Tang, F. Liu, and S. Yan. Body surface context: A new robust feature for action recognition from depth videos. IEEE transactions on circuits and systems for video technology, 24(6):952–964, 2014.

[235] I. Sotgiu. Psychology of emotion: Interpersonal, experiential, and cognitive approaches/paula marie niedenthal, silvia krauth-gruber, françois ric. 2007.

[236] M. Stauffer, P. Maergner, A. Fischer, R. Ingold, and K. Riesen. Offline signature verification using structural dynamic time warping. In 2019 International Conference on Document Analysis and Recognition (ICDAR), pages 1117–1124. IEEE, 2019.

[237] R. Stiefelhagen, C. Fugen, R. Gieselmann, H. Holzapfel, K. Nickel, and A. Waibel. Natural human-robot interaction using speech, head pose and gestures. In 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566), volume 3, pages 2422–2427. IEEE, 2004.

[238] F. Sun, Y. Gu, Y. Cao, Q. Lu, Y. Bai, L. Li, M. Hao, C. Qu, S. Wang, L. Liu, et al. Novel flexible pressure sensor combining with dynamic-time-warping algorithm for handwriting identification. Sensors and Actuators A: Physical, 293:70–76, 2019.

[239] Z. Sun, X. Guo, W. Li, and Z. Liu. Cooperative warp of two discriminative features for skeleton based action recognition. In Journal of Physics: Conference Series, volume 1187, page 042027. IOP Publishing, 2019.

[240] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In International conference on machine learning, pages 1139–1147, 2013.

[241] N. C. Tamer, O. Özdemir, M. Saralar, and L. Akarun. Dynamic time warping based sign retrieval. In 2019 27th Signal Processing and Communications Applications Conference (SIU), pages 1–4. IEEE, 2019.

[242] T. Tamura. Wearable inertial sensors and their applications. In Wearable Sensors, pages 85–104. Elsevier, 2014.

[243] A. B. Tanfous, H. Drira, and B. B. Amor. Sparse coding of shape trajectories for facial expression and action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.

[244] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou. Deep progressive reinforcement learning for skeleton-based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5323–5332, 2018.

[245] M. Tavakol and R. Dennick. Making sense of cronbach's alpha. International journal of medical education, 2:53, 2011.

[246] A. K. Tehrani, M. A. Aghbolaghi, and S. Kasaei. Skeleton-based human action recognition. 2017.

[247] B. Thompson. Foundations of behavioral statistics: An insight-based approach. Guilford Press, 2006.

[248] Y. Tian, L. Cao, Z. Liu, and Z. Zhang. Hierarchical filtered motion for action recognition in crowded videos. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 42 (3):313–323, 2011.

[249] L. Torresani, P. Hackney, and C. Bregler. Learning motion style synthesis from perceptual observations. In Advances in neural information processing systems, pages 1393–1400, 2007.

[250] A. Truong and T. Zaharia. Dynamic gesture recognition with laban movement analysis and hidden markov models. In Proceedings of the 33rd Computer Graphics International, pages 21–24. 2016.

[251] A. Truong, H. Boujut, and T. Zaharia. Laban descriptors for gesture recognition and emotional analysis. The visual computer, 32(1):83–98, 2016.

[252] C.-C. Tseng, J.-C. Chen, C.-H. Fang, and J.-J. J. Lien. Human action recognition based on graph-embedded spatio-temporal subspace. Pattern Recognition, 45(10):3611–3624, 2012.

[253] K. Tumer and J. Ghosh. Error correlation and error reduction in ensemble classifiers. Connection science, 8(3-4):385–404, 1996.

[254] Z. Uzunova, D. Chotrov, and S. Maleshkov. Virtual reality system for motion capture analysis and visualization for folk dance training. In Proceedings of the 12th Annual International Conference on Computer Science and Education in Computer Science (CSECS 2016), Fulda, Germany, pages 1–2, 2016.

[255] K. Van Deemter, B. Krenn, P. Piwek, M. Klesen, M. Schröder, and S. Baumann. Fully generated scripted dialogue for embodied agents. Artificial intelligence, 172(10):1219–1244, 2008.

[256] V. N. Vapnik. The nature of statistical learning. Theory, 1995.

[257] R. C. Veltkamp. Boundaries through scattered points of unknown density. Graphical Models and Image Processing, 57(6):441–452, 1995.

[258] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3d skeletons as points in a lie group. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 588–595, 2014.

[259] D. Ververidis and C. Kotropoulos. Emotional speech recognition: Resources, features, and methods. Speech communication, 48(9):1162–1181, 2006.

[260] B. Vlasenko, B. Schuller, A. Wendemuth, and G. Rigoll. Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing. In International Conference on Affective Computing and Intelligent Interaction, pages 139–147. Springer, 2007.

[261] T. Vogt and E. André. Improving automatic emotion recognition from speech via gender differentiaion. In LREC, pages 1123–1126, 2006.

[262] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 1290–1297. IEEE, 2012.

[263] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu. Cross-view action modeling, learning and recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2649–2656, 2014.

[264] J. Wang, C. Liu, L. Ding, H. Luo, and B. Song. Multi-stage real-time identification for data stream events with drift feature based on dtw. IEEE Access, 7:89188–89204, 2019.

[265] P. Wang, C. Yuan, W. Hu, B. Li, and Y. Zhang. Graph based skeleton motion representation and similarity measurement for action recognition. In European conference on computer vision, pages 370–385. Springer, 2016.

[266] P. Wang, W. Li, J. Wan, P. Ogunbona, and X. Liu. Cooperative training of deep aggregation networks for rgb-d action recognition. arXiv preprint arXiv:1801.01080, 2017.

[267] P. Wang, W. Li, C. Li, and Y. Hou. Action recognition based on joint trajectory maps with convolutional neural networks. Knowledge-Based Systems, 158:43–53, 2018.

[268] S. Wang, J. Li, T. Cao, H. Wang, P. Tu, and Y. Li. Dance emotion recognition based on laban motion analysis using convolutional neural network and long short-term memory. IEEE Access, 2020.

[269] W. Wang, V. Enescu, and H. Sahli. Adaptive real-time emotion recognition from body movements. ACM Transactions on Interactive Intelligent Systems (TiiS), 5(4):1–21, 2015.

[270] X. Wang, J.-F. Hu, J.-H. Lai, J. Zhang, and W.-S. Zheng. Progressive teacher-student learning for early action prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3556–3565, 2019.

[271] V. Wegner Maus, G. Câmara, M. Appel, and E. Pebesma. dtwsat: Time-weighted dynamic time warping for satellite image time series analysis in r. Journal of Statistical Software, 88(5):1–31, 2019.

[272] P. Wei, H. Sun, and N. Zheng. Learning composite latent structures for 3d human action representation and recognition. IEEE Transactions on Multimedia, 21(9):2195–2208, 2019.

[273] D. Weinland and E. Boyer. Action recognition using exemplar-based embedding. In 2008 IEEE Conference on Computer Vision and Pattern Recognition, pages 1–7. IEEE, 2008.

[274] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. Computer vision and image understanding, 104(2-3):249–257, 2006.

[275] A. W. Whitney. A direct method of nonparametric measurement selection. IEEE Transactions on Computers, 100(9):1100–1103, 1971.

[276] G. Willems, T. Tuytelaars, and L. Van Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In European conference on computer vision, pages 650–663. Springer, 2008.

[277] R. J. Williams and D. Zipser. A learning algorithm for continually running fully recurrent neural networks. Neural computation, 1(2):270–280, 1989.

[278] I. H. Witten and E. Frank. Data mining: practical machine learning tools and techniques with java implementations. Acm Sigmod Record, 31(1):76–77, 2002.

[279] L. Xia, C.-C. Chen, and J. K. Aggarwal. View invariant human action recognition using histograms of 3d joints. In 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pages 20–27. IEEE, 2012.

[280] S. Xia, C. Wang, J. Chai, and J. Hodgins. Realtime style transfer for unlabeled heterogeneous human motion. ACM Transactions on Graphics (TOG), 34(4):1–10, 2015.

[281] R. Xiao, Y. Hou, Z. Guo, C. Li, P. Wang, and W. Li. Self-attention guided deep features for action recognition. In 2019 IEEE International Conference on Multimedia and Expo (ICME), pages 1060–1065. IEEE, 2019.

[282] C. Xie, C. Li, B. Zhang, L. Pan, Q. Ye, and W. Chen. Hierarchical residual stochastic networks for time series recognition. Information Sciences, 471:52–63, 2019.

[283] X. Yan and Y. Luo. Recognizing human actions using a new descriptor based on spatial–temporal interest points and weighted-output classifier. Neurocomputing, 87:51–61, 2012.

[284] F. Yang, S. Sakti, Y. Wu, and S. Nakamura. Make skeleton-based action recognition model smaller, faster and better. arXiv preprint arXiv:1907.09658, 2019.

[285] W. Yang, T. Lyons, H. Ni, C. Schmid, L. Jin, and J. Chang. Leveraging the path signature for skeleton-based human action recognition. arXiv preprint arXiv:1707.03993, 2017.

[286] X. Yang and Y. Tian. Effective 3d action recognition using eigenjoints. Journal of Visual Communication and Image Representation, 25(1):2–11, 2014.

[287] X. Yang and Y. L. Tian. Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In 2012 IEEE computer society conference on computer vision and pattern recognition workshops, pages 14–19. IEEE, 2012.

[288] X. Yang, C. Zhang, and Y. Tian. Recognizing actions using depth motion maps-based histograms of oriented gradients. In Proceedings of the 20th ACM international conference on Multimedia, pages 1057–1060, 2012.

[289] W.-J. Yoon and K.-S. Park. A study of emotion recognition and its applications. In International Conference on Modeling Decisions for Artificial Intelligence, pages 455–462. Springer, 2007.

[290] N. Young. An introduction to Hilbert space. Cambridge university press, 1988.

[291] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th international conference on machine learning (ICML-03), pages 856–863, 2003.

[292] L. Yu, D. Xiong, L. Guo, and J. Wang. A remote quantitative fugl-meyer assessment framework for stroke patients based on wearable sensor networks. Computer methods and programs in biomedicine, 128: 100–110, 2016.

[293] M. E. Yumer and N. J. Mitra. Spectral style transfer for human motion between independent actions. ACM Transactions on Graphics (TOG), 35(4):1–8, 2016.

[294] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras. Two-person interaction detection using body-pose features and multiple instance learning. In 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pages 28–35. IEEE, 2012.

[295] M. Zanfir, M. Leordeanu, and C. Sminchisescu. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In Proceedings of the IEEE international conference on computer vision, pages 2752–2759, 2013.

[296] E. Zhang, W. Chen, Z. Zhang, and Y. Zhang. Local surface geometric feature for 3d human action recognition. Neurocomputing, 208:281–289, 2016.

[297] S. Zhang, Y. Yang, J. Xiao, X. Liu, Y. Yang, D. Xie, and Y. Zhuang. Fusing geometric features for skeleton-based action recognition using multilayer lstm networks. IEEE Transactions on Multimedia, 20 (9):2330–2343, 2018.

[298] X. Zhang and C.-W. Shu. On positivity-preserving high order discontinuous galerkin schemes for compressible euler equations on rectangular meshes. Journal of Computational Physics, 229(23):8918–8934, 2010.

[299] L. Zhao, Y. Liu, and N. I. Badler. Applying empirical data on upper torso movement to real-time collision-free reach tasks. SAE transactions, pages 2885–2890, 2005.

[300] Q. Zhao, S. Sun, X. Ji, L. Wang, and J. Cheng. View invariant human action recognition using 3d geometric features. In International Conference on Intelligent Robotics and Applications, pages 564–575. Springer, 2019.

[301] Y. Zhu, W. Chen, and G. Guo. Fusing spatiotemporal features and joints for 3d action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 486–491, 2013.

[302] Q. Zou, W. Tan, E. S. Kim, and G. E. Loeb. Single-and triaxis piezoelectric-bimorph accelerometers. Journal of Microelectromechanical Systems, 17(1):45–57, 2008.

université
PARIS-SACLAY

Titre: INTERACTION BILATÉRALE ENTRE LES ROBOTS HUMANOÏDES ET L'HOMME

Mots clés: Descripteurs basés sur le modèle de Laban, Reconnaissance de gestes, Émotions à partir de gestes corporels, Interaction homme-robot.

Résumé: Dans cette thèse, nous abordons le problème de la reconnaissance du langage corporel humain afin d'établir une interaction bilatérale entre les humains et les robots humanoïde et nous en apportons de nouvelles contributions. Notre approche est fondée sur l'identification de gestes humains, en utilisant une méthode d'analyse de mouvement qui décrit avec précision les mouvements. Cette thèse est constituée de deux parties: la reconnaissance des gestes et la reconnaissance des émotions induites par les gestes. Dans chacun des deux parties, nous mettons en œuvre des méthodes d'apprentissage classiques d'une part et, d'autre part, des méthodes utilisant l'apprentissage profond. Dans la première partie de ce travail, nous avons d'abord défini un descripteur local basé sur l'analyse des mouvements de Laban (LMA), afin de décrire les mouvements. LMA est une méthode permettant de caractériser un mouvement en utilisant quatre composants: Corps, Espace, Forme et Effort. Comme le seul but de cette partie est la reconnaissance gestuelle, seuls les trois premiers facteurs ont été utilisés. L'algorithme Dynamic Time Warping (DTW) est implémenté pour trouver les similitudes des courbes obtenues à partir des vecteurs descripteurs issus de la méthode LMA. Enfin, l'algorithme Support Vector Machine (SVM) est utilisé pour catégoriser les données obtenues. Grâce à la normalisation, notre système est invariant aux positions et orientations initiales des sujets. grâce à l'utilisation des Splines, les données sont échantillonnées afin de réduire la taille des descripteurs et d'adapter les données aux méthodes de classification. Plusieurs expériences utilisant des bases de données publiques ont permis de valider nos choix.
Dans un deuxième temps, nous avons construit un nouveau descripteur basé sur les coordonnées géométriques des différentes parties du corps pour présenter un mouvement. Pour ce faire, en plus des distances entre le centre de la hanche et les autres articulations du corps et des changements angulaires dans le temps, nous définissons les triangles formés par les différentes parties du corps et calculons leur aire. Nous calculons également la superficie de l'enveloppe convexe englobant l'ensemble des articulations. À la fin, nous ajoutons la vitesse des différentes articulations dans le descripteur proposé. Nous avons utilisé un réseau de mémoire à long terme (LSTM) pour évaluer ce descripteur. L'algorithme proposé est mis en œuvre sur deux ensembles de données publiques, NTU RGB+D 120 et SYSU 3D HOI, et les résultats sont comparés favorablement avec ceux disponibles dans la littérature.
Dans la deuxième partie de cette thèse, nous présentons d'abord un algorithme de haut niveau pour identifier les émotions par l'observation des mouvements corporels. Afin de définir un descripteur robuste, deux méthodes sont mises en œuvre : la première est la méthode LMA, complétée du facteur "Effort" alors que la seconde utilise un ensemble de caractéristiques spatio-temporelles. Un pipeline de reconnaissance des mouvements expressifs est proposé afin de reconnaître les émotions des personnes à travers leurs gestes en utilisant des méthodes d'apprentissage automatique (Random Decision Forest, Feed Forward Neural Network). Une étude comparative est faite entre ces deux méthodes afin d'en choisir la meilleure. Notre démarche est validée dans un premier temps grâce à des bases de données publiques, puis par la base de données Expressive Motion (XEM) de gestes expressifs, que nous avons créée à partir de notre propre ensemble de données de gestes expressifs issues du capteur XSENS. Enfin, en appuie de XEM, nous décrivons une étude statistique basée sur la perception humaine afin d'évaluer le système de reconnaissance ainsi que le descripteur proposé. Cela nous permet d'estimer la capacité de notre système à classer et à analyser les émotions comme un être humain. Dans cette partie, deux tâches sont effectuées avec les deux classifieurs (le RDF pour l'apprentissage et l'approche humaine pour la validation).

Title: BILATERAL INTERACTION BETWEEN HUMANOID ROBOTS AND HUMAN

Keywords: Descriptors based on LMA, Gesture Recognition, Emotion based on body gestures, Human-robot interaction.

Abstract: In this thesis, we address the issue of recognizing human body language in order to establish a bi-lateral interaction human-robot and robot-robot. New contributions have been made to this research. Our approach is founded on the identification of human gestures based on a motion analysis method that accurately describes motions. This thesis is divided into two parts: gesture recognition and emotion recognition based on the body gestures. In these two parts, we utilize two methods : classical Machine Learning and Deep Learning.
In the Gesture Recognition section, we first define a local descriptor based on the Laban Movement Analysis (LMA) to describe the movements. LMA is a method that uses four components to describe a movement: Body, Space, Shape and Effort. Since the only goal in this part is gesture recognition, only the first three factors are utilized. The Dynamic Time Warping (DTW) algorithm is implemented to find the similarities of the curves obtained from the descriptor vectors obtained by the LMA method. Finally, the Support Vector Machine, SVM, algorithm is utilized to train and classify the data. Thanks to normalization process, our system is invariant to the initial positions and orientations of people. By the use of Spline functions, the data are sampled in order to reduce the size of our descriptor and also to adapt the data to the classification methods. Several experiments are performed using public data sets.
In the second part of first section, we construct a new descriptor based on the geometric coordinates of different parts of the body in order to characterize a movement. To do this, in addition to the distances between hip center and other joints of the body and the angular changes, we define the triangles formed by the different parts of the body and calculated their area. We also calculate the area of the convex hell encompassing all the joints of the body. At the end we add the velocity of different joints in the proposed descriptor. We used a long short-term memory (LSTM) network to evaluate this descriptor. The proposed algorithm is implemented on two public data sets, NTU RGB+D 120 and SYSU 3D HOI data sets, and the results are compared with those available in the literature.

In the second section of this thesis, we first present a higher level algorithm to identify the inner feelings of human beings by observing their body movements. In order to define a robust descriptor, two methods are carried out: the first method is the LMA with the "Effort" factor, which describes a movement and the state in which it was performed. The second one is based on a set of spatio-temporal features. In the continuation of this section, a pipeline of expressive motions recognition is proposed in order to classify the emotions of people through their gestures by the use of machine learning methods (Random Decision Forest, Feed forward Neural Network). A comparative study is made between these two methods in order to choose the best one. The approach is validated with public data sets and our own data set of expressive gestures called Xsens Expressive Motion (XEM).
In a second part of this section, we carry out of a statistical study based on human perception in order to evaluate the recognition system as well as the proposed motion descriptor. This allows us to estimate the capacity of our system to be able to classify and analyze human emotions. In this part two tasks are carried out with the two classifiers (the RDF for learning and the human approach for validation).

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France