UNIVERSITÀ POLITECNICA DELLE MARCHE
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE DELL'INGEGNERIA
CURRICULUM IN INGEGNERIA INFORMATICA GESTIONALE E AUTOMAZIONE

Applied Machine Learning for Health Informatics: Human Motion Analysis and Affective Computing Application

Ph.D. Dissertation of:
Luca Romeo

Advisor:
Prof. Giuseppe Orlando

Coadvisor:
Prof. Massimiliano Pontil

Curriculum Supervisor:
Prof. Francesco Piazza

XVI edition - new series

UNIVERSITÀ POLITECNICA DELLE MARCHE
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE DELL'INGEGNERIA
CURRICULUM IN INGEGNERIA INFORMATICA GESTIONALE E AUTOMAZIONE

Applied Machine Learning for Health Informatics: Human Motion Analysis and Affective Computing Application

Ph.D. Dissertation of:
Luca Romeo

Advisor:
Prof. Giuseppe Orlando

Coadvisor:
Prof. Massimiliano Pontil

Curriculum Supervisor:
Prof. Francesco Piazza

XVI edition - new series

UNIVERSITÀ POLITECNICA DELLE MARCHE
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE DELL'INGEGNERIA
FACOLTÀ DI INGEGNERIA
Via Brecce Bianche – 60131 Ancona (AN), Italy

alla mia famiglia e ad Anastasia

"Nothing in life is to be feared
it is only to be understood."
Marie Curie

"With all things being equal,
the simplest solution
tends to be the right one."
William of Ockham

Ringraziamenti

Il lavoro svolto ed i risultati raggiunti non sarebbero stati tali senza il supporto di numerose persone che ho avuto modo di conoscere e con cui ho avuto modo di condividere questo percorso di dottorato. Vorrei ringraziare prima di tutto il mio advisor Prof. Giuseppe Orlando ed il Prof. Sauro Longhi per avermi dato la possibilià di svolgere questo percorso di Dottorato, e la Prof. Federica Verdini per aver supervisionato la mia attività di ricerca. Poi voglio esprimere un affettuoso grazie ad i miei colleghi ed amici che in questi anni mi hanno supportato ed hanno contribuito ad accrescere e migliorare il mio lato lavorativo ed umano. Un particolare ringraziamento va sicuramente a Francesco, Sabrina, Davide, Luca, Lucio, Daniele e Marina colleghi ma ancora di più compagni e veri amici.

La mia più sincera gratitudine va anche al mio co-advisor Prof. Massimiliano Pontil, al Prof. Andrea Cavallo e alla Prof.ssa Cristina Becchio che mi hanno dato la possibilità di lavorare e collaborare con il loro gruppo di ricerca presso l'Istituto Italiano di Tecnologia di Genova.

Un grazie anche ai revisori esterni, Prof. Paolo Frasconi ed alla Prof. Nadia Berthouze per i loro importanti suggerimenti al fine di migliorare la tesi.

Infine un ultimo grazie a tutti quelli che hanno creduto in me in questi anni incoraggiandomi e supportandomi sia nelle vittorie che nelle sconfitte.

Ancona, October 2017

Luca Romeo

Abstract

The monitoring of the quality of life and the subject's well-being represent an open challenge in the healthcare scenario. The emergence of solving this task in the new era of Artificial Intelligence leads to the application of methods in the machine learning field.

The objectives and the contributions of this thesis reflect the research activities performed on the topics of (i) human motion analysis: the automatic monitoring and assessment of human movement during physical rehabilitation and (ii) affective computing: the inferring of the affective state of the subject.

In the first topic, the author presents an algorithm able to extract clinically relevant motion features from the RGB-D visual skeleton joints input and provide a related score about subject's performance. The proposed approach is respectively based on rules derived by clinician suggestions and machine learning algorithm (i.e., Hidden Semi Markov Model). The reliability of the proposed approach is tested over a dataset collected by the author and with respect to a gold standard algorithm and with respect to the clinical assessment. The results support the use of the proposed methodology for quantitatively assessing motor performance during a physical rehabilitation.

In the second topic, the author proposes the application of a Multiple Instance Learning (MIL) framework for learning emotional response in presence of continuous and ambiguous labels. This is often the case with affective response to external stimuli (e.g., multimedia interaction). The reliability of the MIL approach is investigated over a benchmark database and one dataset closer to real-world problematic collected by the author. The obtained results point out how the applied methodology is consistent for predicting the human affective response.

Sommario

Il monitoraggio della qualità della vita e del benessere della persona rappresenta una sfida aperta nello scenario sanitario. La necessità di risolvere questo task nella nuova era dell'Intelligenza Artificiale porta all'applicazione di metodi dal campo del machine learning.

Gli obiettivi e i contributi di questa tesi riflettono le attività di ricerca svolte (i) nell'ambito dell'analisi del movimento: valutazione e monitoraggio automatico del movimento umano durante la riabilitazione fisica, e (ii) nell'ambito dell'affective computing: stima dello stato affettivo del soggetto.

Nel primo tema il candidato presenta un algoritmo in grado di estrarre le caratteristiche di movimento clinicamente rilevanti dalle traiettorie dello skeleton acquisite da un sensore RGB-D, e fornire un punteggio sulla prestazione del soggetto. L'approccio proposto si basa su regole derivate da indicazioni cliniche e su un algoritmo di machine learning (i.e., Hidden Semi-Markov Model). L'affidabilità dell'approccio proposto è studiata su un dataset collezionato dal candidato rispetto ad un algoritmo gold standard e alla valutazione clinica. I risultati sostengono l'uso della metodologia proposta per la valutazione quantitativa delle prestazioni motorie durante la riabilitazione fisica.

Nel secondo topic il candidato propone l'applicazione del framework di Multiple Instance Learning per l'apprendimento della risposta emotiva in presenza di label continui ed ambigui. Questa varaibilità è spesso presente nella risposta affettiva ad uno stimolo esterno (e.g., interazione multimediale). L'affidabilità dell'approccio di Multiple Instance Learning è indagata su un database di benchmark e un dataset più vicino alle problematiche del mondo reale acquisito dal candidato. I risultati ottenuti evidenziano come la metodologia proposta è consistente per la stima dello stato affettivo.

Acronyms

ML Machine Learning

AI Artificial Intelligence

HI Health Informatics

HCI Human Computer Interaction

ICT Information and Communication Technologies

IT Information Technology

CE Consumer Electronics

RGB Red Green Blue

RGB-D Red Green Blue - Depth

DF Dynamic Features

SF Static Features

AE Absolute Error

RE Relative Error

RMSE Root Mean Square Error

TFs Target Features

TVFs Target Velocity Features

PFs Postural Features

HMM Hidden Markov Model

HSMM Hidden Semi Markov Model

DTW Dynamic Time Warping

CS Clinical Score

AC Affective Computing

CBR Content-Based Recommender

SAM Self-Assessment Manikin

GSR Galvanic Skin Response

HR Heart rate

HRV Heart rate variability

RR Inter-beat interval

ECG Electrocardiogram

BP Blood Pressure

OXY Blood Oxygen Saturation

RESP Respiration

ST Skin Temperature

EMG Electromyography

Accel Accelerometer

PHO Mobile phone

SVM Support Vector Machine

NN Neural Networks

QDA Quadratic Discriminant Analysis

K-NN K-Nearest Neighbors

RF Random Forest

LSD Local Scaling Dimension

ACC Accuracy

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

MIL Multiple Instance Learning

DD Diverse Density

NB Naive Bayes

ROC Receiver Operating Characteristic

AUC Area Under Curve

MTL Multi-Task Learning

MKL Multiple Kernel Learning

Contents

List of Figures

List of Tables

Chapter 1.

Introduction

"The time is ripe for our measurement system to shift emphasis from measuring economic production to measuring people's well-being."
The Stiglitz Report [1]

With the ubiquitous presence of biomedical data and its increasing importance in a wide range of healthcare applications such as Computational Biology, Clinical Informatics, Rehabilitation, and Psychology, there is a growing demand for automated and/or semi-automated analysis for accelerating basic science discoveries and facilitating evidence-based clinical solutions. Machine Learning (ML) offers a realm of possibilities for discovering meaningful pattern or structures of data using optimization and probabilistic methods. In a broad sense, there are two types of applications in biomedical informatics where ML is commonly used: (i) the knowledge discovery by analyzing historical data to go insights on what happened and why it happened and (ii) the design of a decision-making application, building a predictive model and scale it to make predictions using unseen data.

The monitoring of the quality of life and the subject's well-being represents an open challenge in this scenario. The World Health Organization [2] defined health as "a state of complete physical, mental and social well-being" [3]. This statement emphasizes the importance of emotional well-being for health and is supported by an increasing body of epidemiological, social science, experimental research that is beginning to suggest that initiatives which aim to promote physical well-being excluding mental and social well-being may be doomed to failure. However, the monitoring and the estimation of subjects well-being and the measure of the quality of life remain an open question. The emergence of solving this task in the new era of Artificial Intelligence (AI) leads to the design and building of new methods in computer science and machine learning scenario. In particular, the "Prioritizing of Human Well-being in the Age of Artificial Intelligence" is the title of the Report of IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems' [4]. Recently, Fabrice Murtin, Senior Economist, Household Statistics and Progress Measurement Division of the OECD Statistics Directorate pointed out how the changes that are brought forward by these digital AI technologies could be monitored across

all the various dimensions of human well-being. This represents a possibility to close the loop: the AI technologies and ML methods can improve the well-being and the quality of life measuring at the same time the subject's well-being during human-computer or human-robot interaction.

The objectives and the contributions of this thesis reflect the research activities performed on the following two main topics:

- Design and implementation of an algorithm for **monitoring the physical well-being** of subjects: automatic evaluation and assessment of human movement during physical rehabilitation.
- Design and implementation of an algorithm for **monitoring the emotional well-being** of subjects: inferring the affective state of the user during multimedia interaction.

There is much to be learned through the discovery of new ML methodologies able to solve this two problems and at the same time promising a suitable design, implementation, interpretation, and validation of these methods from a computer scientist, medical and humanistic perspective. Hence, this thesis is titled *Applied Machine Learning for Health Informatics: Human Motion Analysis and Affective Computing Application*. The remainder of this chapter is organized as follows:

- Section 1.1 provides a broad background and motivation for this thesis topic, author motivates the research study and highlights its significances from the perspectives of Health Informatics (HI), Artificial Intelligence (AI), and Human-Computer Interaction (HCI).
- Section 1.2 presents the thesis statement, where the problem is formally defined from a machine-learning point of view with a list of specific research questions answered in this thesis.
- Section 1.3 resumes the index of the thesis.
- Section 1.4 presents the thesis outcomes in terms of scientific publications.

1.1. Background and Motivation

Health Informatics is concerned with the use of computational intelligence for the management of processes relevant for human health and well-being, ranging from the collective to the individual. In healthcare, machine learning could help providing more accurate diagnoses and more effective healthcare services, through advanced analysis that improves decision-making', according to a report on AI by The Royal Society [5]. Recently, Deep Learning and other advanced machine learning technologies have

revolutionized in computer vision, speech recognition, and natural language processing and brought promising results in many other areas. Despite this, applying these AI revolutions to human health and wellness problems remains some challenges [6]. For instance, the following five methodological and technical challenges should be faced:

1. **Representation of subjective knowledge**: much of knowledge in well-being science is subjective. For instance, the subjects' personality and the subjects' mood should be represented with concrete and consistent mathematical structures in order to be embedded in a ML model.

2. **Understand things that humans do not**: right now machine learning research is interested in getting computers to be able to understand data that humans do: images, text, sounds, and so on. However, the focus in novel advanced machine learning techniques such as deep learning is to understand things that humans do not.

3. **Data analysis issues**: biomedical data analysis deals with different challenges such as large volumes of data, high dimensions, imbalanced classes, heterogeneous sources, noisy data, incompleteness, rich contexts, weakly structured or unstructured data, noisy and ambiguous labeling. Optimization and ML algorithms are developed to face these type of problems. It is also of much interest to study and revisit traditional machine-learning topics such as clustering, classification, regression, and dimension reduction and turn them into powerful customized approaches in order to solve these challenges.

4. **Models, Reasoning, and Inference, Data interpretation**: the reasoning about data through representations should be understandable to the human being. For instance, the goal of ML is not only to increase accuracy rate of predictions but also understand the causality with reliable models, reasoning, and inference. The sheer volume and complexity of the data that is possible to acquire nowadays in biomedical informatics present major barriers toward their translation into effective clinical meaning and actions.

5. **User-centered design system**: it is important to understand how the AI revolution affects subjects' emotions and their quality of life and how to design a human-centered system.

1.1.1. The Challenge of Evaluating Human Movement

Global population is aging rapidly and life expectancy is constantly growing. According to the Global Health Observatory data, 71.4 years was the average life expectancy at birth of the global population in 2015. In this scenario living longer does not necessarily mean living healthier, more active and independent. Active and healthy aging is thus one of the most important societal challenges shared by all governments,

researchers and industries of different areas [2] to achieve the health Sustainable Development Goals [7]. In this scenario, it is not surprising that eHealth and digital health are among the fastest-growing areas in the Consumer Electronics (CE) and Information Technology (IT) markets. Increased acceptability and usability, ubiquitous access to data, real-time clinical outcomes, remote patient-to-physician experiences in telemedicine and telerehabilitation are only some of the (r)evolutions coming in the latest years. In stark contrast to such possibilities, the technology itself is certainly easily accessible and cost-effective. In this scenario, telerehabilitation is a solution for delivering services at home, supporting patients and clinicians by minimizing the barriers of distance, time and cost. Although telerehabilitation platforms based on vision and wearable sensors are widely spread [8, 9, 10], the way to ensure a continuous monitoring of body motion and an accurate evaluation of rehabilitation therapy remains a challenge. The primary challenge is to design and realize an Information and Communication Technology (ICT) rehabilitation tool at home on the basis of acceptability, low cost and connectivity principles. Many recent works reported in the scientific literature focused on the design of such systems by using CE devices and, sometimes, AI techniques. Preventive healthcare systems have been proposed by using wireless sensor networks [11] or consumer home networks [12]. Accordingly, some studies address the issue of human activity recognition in a domestic scenario focusing on the automatic visual detection and recognition of human behavior [13, 14], including abnormal and dangerous activity detection [15], normal motion patterns [16], daily [17, 18, 19] and sport activities [20, 21, 22, 23]. On the same time different eHealth software platforms have been proposed, as in [24] where authors aim to monitor patients physiological data through a Smart TV and an Open Services Gateway initiative (OSGi) architecture and as in [25] where a wireless blood pressure measurement device is designed and works in cooperation with a web-based platform. One of the most promising aspects of among all the eHealth applications is the telerehabilitation. The primary driver behind telerehabilitation is the need to eliminate the inequality of access to rehabilitation services. In addition to CE and ICT companies, many other stakeholders are involved in telerehabilitation systems, including patients, physicians, engineers, and, obviously, academic researchers. A recent trend in the design of such multidisciplinary systems is the so-called "collaborative design" paradigm, a process involving all these actors [26]. Patient-centered approaches may help to meet patients' needs increasing acceptability and usability, while physicians are involved in the definition of inputs and outputs of the ICT platform (e.g. targets, assessment, scoring). From an engineering perspective, a telerehabilitation system is based on a signal processing stage (automated segmentation of images [27] or biosignal analysis), identification [28] and assessment of movements (employing machine learning and/or action similarity algorithms). However one of the limits in the current scenario is the (almost) complete lack of involvement of physicians in the algorithmic core of the system. While on one side this can seem a normal situation, on the other it drasti-

cally decreases the usefulness and interpretability of the outcomes of the entire system thus limiting the advantages of the "collaborative design" paradigm (**data interpretation issue**). The telerehabilitation tool should provide a functional monitoring of the motion during exercise execution, such as a physiotherapist does during the ambulatory training. The system should at the same time understand the causality of motion data with reliable models, reasoning, and inference providing a suitable feedback for supporting both clinicians and patients during the rehabilitation process. Accordingly, the data analysis procedure should deal with different problems such as (**data analysis issues**):

- noise: sensor data is subject to several sources of errors such as hardware noise, interference, and noise from external sources and environment, inaccuracies, and imprecision. That means the low-cost motion tracking system should be validated in the rehabilitation scenario.

- high dimensional data: often the human motion tracking data has high dimensions. Hence, a suitable features extraction stage or dimensionality reduction and features selection techniques should be exploited in this context. In addition, also some ML algorithms (i.e., Support Vector Machines, Sparse Support Vector Machine) are designed to implicitly handle high dimensional data.

- incompleteness and missing data: the algorithm should manage situations of occlusion during the acquisition stage. For instance, the visual sensors are widely used for human motion tracking. However, the rehabilitation exercise monitoring involves considering dynamic movements with a wide range of motion and issues related to the joints tracking such as joints occlusion due to postures adverse to the vision sensor.

The accuracy and precision of the motion evaluation algorithm in a telerehabilitation system should be comparable with respect to the clinician analysis. The algorithm should identify which body segments are making a mistake and monitor simultaneously multiple motion features with high accuracy, otherwise captured by the human eye with difficulty (**Understand things that humans do not**).

1.1.2. The Challenge of Inferring Human Emotion

Emotion is a psycho-physiological response triggered by conscious and/or unconscious stimuli that cannot be explained by scientific principles such as rational thought, logical arguments, testable hypotheses, and repeatable experiments. Emotions play a crucial role in human communication and can be expressed by multidimensional cues such as vocabulary, the intonation of voice, facial expressions and gestures.

The term affect refers to emotion and related phenomena such as:

- Emotions (e.g., angry, sad, joyful, fearful)
- Moods (e.g., cheerful, gloomy, irritable, listless, depressed, buoyant)
- Interpersonal stances (e.g., distant, cold, warm, supportive)
- Preferences/Attitudes/Sentiment (e.g., liking, loving, hating)
- Personality (e.g., nervous, anxious, reckless, morose)
- Culture (e.g., Individualistic vs. Collectivist; engineering vs. social sciences)

The relation between emotion and computer science lays hold of various of computer science such as HCI, AI, and HI.

Most of the current HCI systems are unable to identify human emotional states and use this information in the decision-making process. The importance of affect for HCI can be explained by denoting its effects on three cognitive processes:

- **Attention**: affective processes have a way of being completely absorbing and to capture attention. Basically, they direct and focus our attention on those objects and situations that have been appraised as important to human's needs and goals [29].
- **Memory**: affect has also an implication for learning and memory [30, 31]. Events with an affective load are generally remembered better than events without such a load, with negative events being dominant over positive events [32].
- **Decision making**: affective processes also have their influence on our flexibility and efficiency of thinking and problem-solving [33]. It has also been demonstrated that affect can (heavily) influence judgment and decision making [34, 35].

Then, the main goal of HCI research is to create interfaces that are both efficient and effective as well as enjoyable and satisfying [36].

In 1935, Flanders Dunbar noted that the *"Scientific study of emotion and of the bodily changes that accompany diverse emotional experience marks a new era in medicine"* [37]. Then the emotions are now being given a position in health informatics and for application such as the support/assistance of independent living, chronic disease management, facilitation of social support, and to decrease the barriers of distance, time and cost. Many physiological processes that are profound significance for health can be influenced by way of emotions [38, 39, 40]. Accordingly, emotions play an important role in chronic disease, cancer, and rehabilitation process. There is a strict connection between emotion and stress. In particular, stress is a physiological

response to a mental effort, emotional or physical. It can be defined as the reaction of a person to environmental influences or physical demand. Stress condition can affect the physical and emotional wellbeing leading symptoms such as headaches, stomachaches and sleeplessness, and insomnia. In recent years the impact of stress on society has increased. A study conducted by the American Institute of Stress [41] disclosed how in 2015, the 48% of people believe that their stress condition has increased over the last five years, and 77% of people regularly experience physical symptoms caused by stress with a negative impact on personal and professional life. Accordingly, the influence of stress and its consequences on society also affects the economic aspect; causing a loss of productivity, increasing health care costs and social welfare costs [42]. Hence, emotions, mood, and stress can affect not only temperament, personality, disposition, and motivation but also the person's physical well-being, judgment, and perception. Since emotions are complex and move in various directions the first step towards emotion analysis is the identification and categorization of different emotions [43]. Modeling emotional feelings and considering their behavioral implications are useful in preventing emotions from having a negative effect on the workplace. Accordingly, the decision-making process should discard emotion whenever possible: both positive and negative emotions can distort the validity of a decision.

Almost half a century ago, the American psychologist Ulric Neisser stated [44] that "Human thinking begins in an intimate association with emotions and feelings which is never entirely lost". From the Publication of Picard's book [45], AI starts to place the emotion into account for understanding human cognition. The term affective computing was originally defined in [45] as computing that relates to, arises from, or influences emotions. Hence, affective computing assumes that there is a benefit to give computers "emotional intelligence". This presents unique challenges and several opportunities for signal processing and machine learning researchers to solve the complex task of detecting emotional cues occurring during HCI and subsequently synthesizing emotional response. Affective computing is an interdisciplinary field of research that covers but is not limited to the topics involving: sensing and analysis (i.e., recognition of human emotion), psychology and behavior analysis, behavior generation and user interaction. In particular some research topics about affective computing research include [46]:

Emotion Recognition in:

- Speech:
  - Emotion in natural speech
  - Depression detection
- Text
  - Opinions in facebook, twitter, instagram; blogs
  - Emoticons

- Face
  - Understanding impact of aging
  - Recognizing expressions with thermal, RGB-D image
- Physiology
  - Detecting stress from skin conductance
  - Inferring emotional response during multimedia interaction

Synthesis

- Emotional speech
- Emotional facial expressions

Modeling

- Modeling emotional influences on decision making
- Modeling factors that elicit emotions

Applications

- Health: Detection and shaping
- Education: Detection and shaping
- Behavioral science
- Games/Entertainment computing
  - Responses to victory and defeat
  - Affective music player
  - Boredom Detection
- Automotive

The recognition of emotion and the estimation of stress level disclose several challenges. In particular, the **data analysis** procedure should be robust to:

- high dimension data. For instance, data acquired from Electroencephalogram and EMG lead to high-dimensional data sets with much more features than data items. Also here, features selection, dimensionality reduction, and suitable ML algorithm should deal with this issue.

- imbalanced class. Often the data acquisition procedure in the emotion/stress recognition task leads to an unbalanced class, where the samples are not uniformly distributed over all labels. In this case, it is often preferable to maximize other metrics such as macro/micro F1 score instead of accuracy. Moreover, data level approach focuses on increasing the frequency of the minority class or decreasing the frequency of the majority class (resampling techniques). While the classification level approach aims to modify existing classification method to make them appropriate for imbalanced data sets (e.g., ensemble classifiers).

- heterogeneous sources. Most of the affective computing studies monitor the user with several sensors or sensor channels, which can be considered as such co-occurring sets. The model prediction should combine several sensors and at the same time learns the importance of individual sensors. For instance, a technique such as Multiple View Learning (MVL) could be used for combining different sources and learning a joint model over all sensor data.

- ambiguous label. The ground truth label cannot always be reliable, reflecting ambiguous or summative global emotional responses. Accordingly, the emotion is time-varying and cannot consistent with all the observation sequence.

- noise: sensors for physiological data acquisition is subject to several interferences including also artifact noise. A preliminary consistent pre-processing stage is required to remove outliers.

Obviously, the emotional response is subjective (**Representation of subjective knowledge**). The ML algorithm should model the emotional response of a single user and at the same time generalize across different users. Hence, the implemented ML model should provide a robust performance for a single user (*user-specific* model) and also across users (*user-independent* model). Often emotions and thoughts are best kept hidden than displayed in front of people. There are many different reasons that human may endeavor to hide or disguise, but what they have in common is that they are all fear-induced. Hence, it is difficult to understand and detect the emotion and the mood state of a subject. In this context, affective computing and machine learning aim to discover also what is hidden by the human being (**Understand things that humans do not**).

1.1.3. The human machine closed-loop model

Figure 1.1 and 1.2 show respectively the human-machine closed-loop model for telerehabilitation and affective computing scenario. These closed loop models take the human/user/subject into the loop and include sensors, processing, modeling, and actuators. Closed loop models can be concisely defined as a control system with an active feedback loop, allowing the control unit to dynamically compensate for changes in

the system. Rehabilitation and affective computing are examples of this closed loop model.

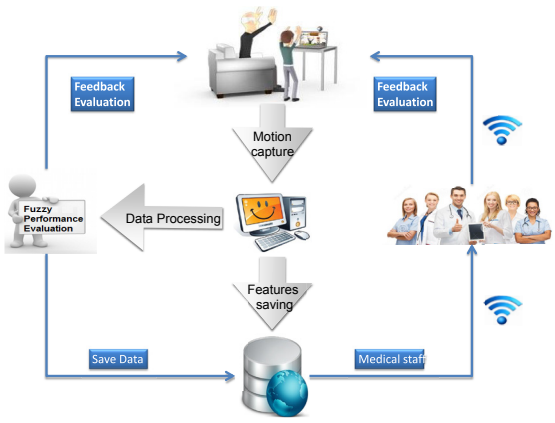


Figure 1.1.: The telerehabilitation framework

In telerehabilitation scenario, the data processing aims to monitor and to assess the human movement. Hence, a properly feedback is sent to patients and clinicians in order to improve the execution performance and track the patients' progress over the time.

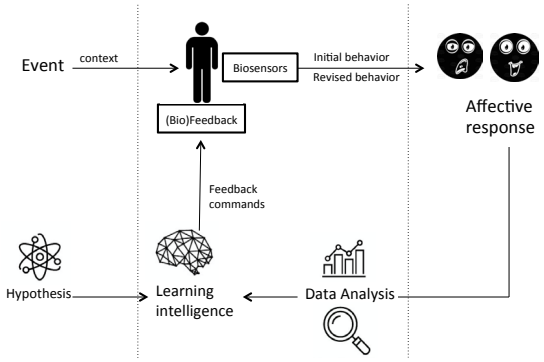


Figure 1.2.: The affective computing model

In the affective computing scenario, the vast majority of research has not applied closed-loop models and focus the attention only on the data analysis and the design of machine learning model able to estimate the affective state of the user. Examples of affective closed loops are for instance a computer/robot/avatar that adapts its interaction dialogue to the level of frustration of its user, or a recommender multimedia system that chooses the music/video to be played so as to guide the user to a better mood, or a teacher that adapts its interactive lesson to the level of interest of the audience.

1.2. Thesis: Problems statement

Thesis statement: With computational techniques, including advanced machine learning algorithms, the main objective is to design and develop an algorithm for monitoring the physical and emotional well-being of a subject.

1.2.1. Problem 1: Quantitative Assessment of Human Motion

Problem 1 is addressed to monitor the physical well-being of the subject: design and develop algorithms for real-time assessing exercise performance during physical rehabilitation stage.

The research questions regarding the machine learning algorithm for the movement assessment stage are summarized below:

1. How can ML model be applied for assessing the human movement with respect to a reference example or a set of rules?

2. How can the algorithm provide a suitable feedback for supporting both clinicians and patients during rehabilitation process?

Questions-related to the evaluation scheme include:

1. How can an objective measurement be designed to validate the proposed algorithm?

2. Does the proposed ML algorithm outperform standard algorithm widely used in literature?

3. How far/close is the proposed algorithm from clinician evaluation of exercise performance?

1.2.2. Problem 2: Emotion Inference from Physiological predictors

Problem 2 is addressed to monitor the emotional well-being inferring the affective state and the level of stress of the subjects.

The research questions regarding the machine learning algorithm for the emotion recognition task are summarized below:

1. How can ML model be applied to infer the affective state of the user and model the variability in physiological response over the course of multimedia interaction?

2. How can ML model be applied to handle the ambiguity and the change over the time of the emotional response?

Questions-related to the evaluation scheme include:

1. Does ML method outperform standard supervised algorithm based on video-level features?

2. Is ML method reliable for the emotion recognition task towards the real world usage?

1.3. Thesis overview

This thesis aims to answer the question reported above designing and developing machine learning algorithms for the monitoring of physical (Quantitative Assessment of Human Motion) and emotional well-being (Emotion Inference from Physiological predictors).

Figure 1.3 together with the following list show the organization and an overview of the rest of the thesis.

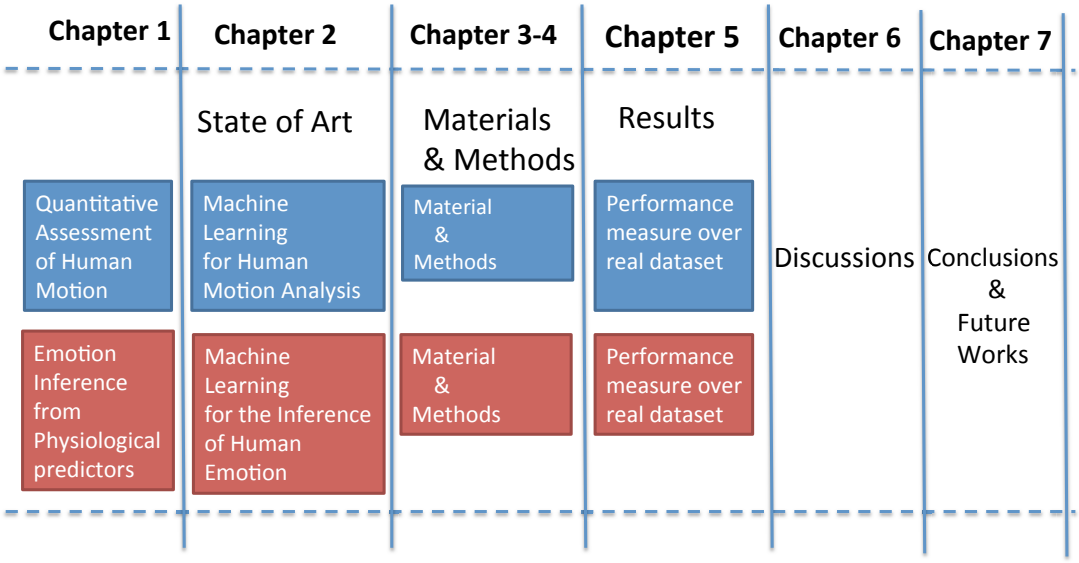


Figure 1.3.: The organization of the rest of the thesis

- **Chapter 2** reviews related literature in two research fields: Human motion assessment and Emotion recognition using Physiological signals. The author outlines the strengths and the drawbacks of the existing achievements, focusing on what problem questions they answer and what they do not.
- **Chapter 3-4** describes the proposed materials and methods. In particular, the author presents the proposed ML algorithms for the data processing/analysis stage.
- **Chapter 5** firstly presents the Experimental Protocol including the adopted dataset, the experimental setup, and measure. Then, the author provides the experimental results for evaluating the performance of the proposed methods.
- **Chapter 6** discusses the obtained results and provides further statistical analyses.
- **Chapter 7** presents the conclusions and future works.

1.4. Thesis outcomes

The detailed descriptions of the thesis outcomes are available in the follow publications. The contribution of the candidate are reported below each paper.

1.4.1. Problem 1: Quantitative Assessment of Human Motion

- Journal
  - M. Capecci, L. Ciabattoni, F. Ferracuti, A.Monteriù, L. Romeo and F. Verdini (in press), *Collaborative design of a telerehabilitation system enabling virtual second opinion based on fuzzy logic*, IET Computer Vision. doi 10.1049/iet-cvi.2017.0114
    L. R. developed the study concept, L.R. performed the kinematic data collection, L.R. implemented the methodology and the web-based platform, L.R. drafted the manuscript providing critical revisions and approved the final version of the manuscript for submission.
  - M. Capecci, M. G. Ceravolo, F. Ferracuti, M. Grugnetti, S. Iarlori, S. Longhi, L. Romeo and F. Verdini (in press), *An instrumental approach for monitoring physical exercises in a visual markerless scenario: a proof of concept*, Journal of Biomechanics.
    doi 10.1016/j.jbiomech.2018.01.008
    L. R. developed the study concept, L.R. performed the kinematic data collection, L.R. implemented the methodology, L.R. planned and performed

the data analysis, L.R. drafted the manuscript providing critical revisions and approved the final version of the manuscript for submission.

- M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, V. Kyrki, A. Monteriù L. Romeo and F. Verdini, *A Hidden Semi-Markov Model based Approach for Rehabilitation Exercise Assessment*, Journal of Biomedical Informatics. Volume 78, 2018, Pages 1-11.
  L. R. developed the study concept, L.R. performed the kinematic data collection, L.R. implemented the methodology, L.R. planned and performed the data analysis, L.R. drafted the manuscript providing critical revisions and approved the final version of the manuscript for submission.

- Conference Proceeding
  - L. Ciabattoni, A. De Cesare, G. Foresi, A. Monteriù, D. Proietti Pagnotta, L. Romeo and L. Spalazzi (accepted), *Complex Activity Recognition System Based on Cascade Classifiers and Wearable Device Data*, IEEE International Conference on Consumer Electronics (ICCE 2018).
    L.R. planned and performed the data analysis, L.R. provided critical revisions and approved the final version of the manuscript for submission.

  - L. Ciabattoni, F. Ferracuti, G. Lazzaro, L. Romeo and F. Verdini, *Serious gaming approach for physical activity monitoring: A visual feedback based on quantitative evaluation*, IEEE 6th International Conference on Consumer Electronics-Berlin (ICCE-Berlin 2016).
    L. R. developed the study concept, L.R. supported the implementation of the methodology, L.R. drafted the manuscript providing critical revisions and approved the final version of the manuscript for submission.

  - M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, S. Longhi, L. Romeo, N. Russi Severino and F. Verdini, *Accuracy evaluation of the Kinect v2 sensor during dynamic movements in a rehabilitation scenario*, 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2016)
    L. R. developed the study concept, L.R. performed the kinematic data collection, L.R. planned and performed the data analysis, L.R. drafted the manuscript providing critical revisions and approved the final version of the manuscript for submission.

  - L. Ciabattoni, F. Ferracuti, S. Iarlori, S. Longhi and L. Romeo, *A novel computer vision based e-rehabilitation system: From gaming to therapy support*, IEEE International Conference on Consumer Electronics (ICCE 2016).

L. R. developed the study concept, L.R. performed the kinematic data collection, L.R. implemented the methodology and the web-based platform, L.R. drafted the manuscript providing critical revisions and approved the final version of the manuscript for submission.

- M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, V. Kyrki, S. Longhi, L. Romeo and F. Verdini, *Physical rehabilitation exercises assessment based on hidden semi-markov model by kinect v2*, International Conference on Biomedical and Health Informatics (BHI 2016).
  L. R. developed the study concept, L.R. performed the kinematic data collection, L.R. implemented the methodology, L.R. planned and performed the data analysis, L.R. drafted the manuscript providing critical revisions and approved the final version of the manuscript for submission.
- M. Capecci, M. G. Ceravolo, F. D'Orazio, F. Ferracuti, S. Iarlori, G. Lazzaro, S. Longhi, L. Romeo M. Grugnetti, S. Iarlori, S. Longhi, L. Romeo and F. Verdini, *A tool for home-based rehabilitation allowing for clinical evaluation in a visual markerless scenario*, 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2015).
  L. R. developed the study concept, L.R. performed the kinematic data collection, L.R. implemented the methodology, L.R. planned and performed the data analysis, L.R. drafted the manuscript providing critical. revisions and approved the final version of the manuscript for submission.

1.4.2. Problem 2: Emotion Inference from Physiological predictors

- Conference Proceeding
  - L. Ciabattoni, E. Frontoni, D. Liciotti, M. Paolanti and L. Romeo (accepted), *A Sensor Fusion Approach for Measuring Emotional Customer Experience in an Intelligent Retail Environment*, IEEE 7th International Conference on Consumer Electronics-Berlin (ICCE-Berlin 2017).
    L. R. developed the study concept, L.R. implemented the methodology, L.R. drafted the manuscript providing critical revisions and approved the final version of the manuscript for submission.
  - L. Ciabattoni, F. Ferracuti, S. Longhi, L. Pepa, L. Romeo and F. Verdini, *Real-time mental stress detection based on smartwatch*, IEEE International Conference on Consumer Electronics (ICCE 2017).
    L. R. developed the study concept, L.R. implemented the methodology, L.R. provided critical revisions and approved the final version of the manuscript for submission.

– L. Ciabattoni, F. Ferracuti, S. Longhi, L. Pepa, L. Romeo and F. Verdini, *Multimedia experience enhancement through affective computing*, IEEE International Conference on Consumer Electronics (ICCE 2017).
L. R. developed the study concept, L.R. implemented the methodology, L.R. provided critical revisions and approved the final version of the manuscript for submission.

Chapter 2.

State of the art

The state of art related to human movement analysis and the methodology for quantitatively assess human movement is summarized in Section 2.1. While the literature reviews of affective computing and the emotion recognition using physiological signals are described in Section 2.2. In Section 2.3 author reports some study within the intersection of physical rehabilitation and affective computing focusing in the possible emotional effects on movement execution in patient with chronic pain.

2.1. State of art: Quantitative Assessment of Human Movement

The vision sensors employed for human movement assessment are shown in Section 2.1.1. The pre-processing and features extraction stage related to this type of data are described in Section 2.1.2. Then, the existing motion assessment algorithms are resumed in Section 2.1.3, divided in rule-based (see Section 2.1.4) and template-based method (see Section 2.1.5). The discussions (see Section 2.1.6) are provided in order to highlight the drawback of the method presented in literature. Finally, the author's contribution is clarified and described in Section 2.1.7.

2.1.1. Sensors

In the last years, many research projects focused on developing affordable, acceptable and reliable telerehabilitation applications, wearable and vision sensors based [47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 8, 57, 58]. For a complete review of the wearable-based system for movement assessment, the reader can refer to [9, 10]. Generally, complex and intrusive technologies, able to accurately monitor human motion as electromyography [59], optoelectronic motion analysis or wearable inertial systems cannot be routinely adopted in a physiotherapy ambulatory or at home, because their costs and low acceptability and usability, as defined by the Unified Theory of Acceptance and Use of Technology criteria (UTAUT [60]). On the other hand, more than one wearable sensor (i.e., accelerometer) is required to accurately describe mo-

tion and posture [10, 61] and to estimate related parameters. This situation disagrees with the acceptance requirement (UTAUT [60]). Therefore, vision-based systems are preferable for monitoring the whole body motion during the execution of a functional movement in a delimited environment.

Vision Sensors

For a complete review of the vision-based systems for movement assessment, the reader can refer to the papers [8, 62]. The visual based tracking systems can be divided in marker-based and marker-free. Although the former allows an accurate tracking of the motion, the latter are no obtrusive and they overcome the mutual occlusion problem allowing to perform the motion analysis in a three-dimensional space. The Red-Green-Blue Depth (RGB-D) camera overcomes limitations of the traditional RGB camera. In particular, the value of each pixel in a depth image indicates the calibrated distance between camera and scene. The depth information is used to address the following issues:

- Reconstruction of the 3D structure of the scene;
- Human body pose estimation and tracking, object recognition and tracking;
- Implementation of Natural User Interface (NUI);
- Solving ambiguities (good invariance against color, texture and illumination changes);
- Image synthesis

Three main sensing RGB-D technologies are applied in computer vision research:

- stereocamera;
- time of flight cameras (ToF);
- structured light sensors

Figure 2.1 shows three examples of stereocamera, ToF, and structured light sensors.

(a) (b) (c)

Figure 2.1.: Bumblebee 2 Sony stereo vision camera (a), Microsoft Kinect (PrimeSense) (b), Microsoft Kinect v2 (c)

Table 2.1 shows the comparison between the three sensing technologies in terms of resolution, speed, range, depth resolution, the field of view, holes in the depth map, price and invariance against illumination.

Table 2.1.: Comparison among depth sensors technologies

Sensor Type	**Stereo camera**	**ToF**	**Structured Light**
Resolution	640 x 480 or more	64 x 48 to 512 x 424	640 x 480
Speed	Slow	Fast	Fast
Range	Only limited by baseline	5 m to 10 m (indoors or outdoors)	0.8-3.5 m (typically indoors)
Depth resolution	Depends on camera baseline and resolution	Less than 5 mm	Less than 1 cm
Field of view	Depends on camera lenses	Approx. 43°(v), 69°(h)	Approx. 43°(v), 57°(h)
Holes in depth map	Yes	No	Yes
Price	Quite cheap	Expensive	Cheap
Invariance against illumination	Yes	No	No

Microsoft Kinect is a cheap, unobtrusive and easy to set up technology that could be usefully applied to monitor subjects during rehabilitation programs. Microsoft Kinect for Windows SDK includes a skeletal tracking where the 20 virtual anatomical joint trajectories are extracted from depth map with a per-pixel semantic segmentation approach based on random decision forest algorithm [63]. In 2014, the second version of Microsoft Kinect (Kinect v2) was released [64]. It employs a novel time-of-flight technology, compared with the previous version that falls within the category of Structured Light camera. For further detail in the depth-sensing technology implemented in the new generation of Kinect v2, the reader can refer to [64]. Compared with the previous version, Kinect v2 provides a superior depth map resolution (512×424 vs 320×240), allowing to recognize thin objects and solving some ambiguity problems. The increase in resolution allows identifying 25 distinct virtual body joints (see Fig. 3.3.a).

In the telerehabilitation scenario, Microsoft Kinect is used at home as unobtrusive and low-cost assistive technology for human action recognition [65, 22, 66], fall detection [67], gait measurement [68] and for supporting patients and physiotherapists in the rehabilitation cycle [49, 50, 69]. It has been integrated into a telerehabilitation system to provide physiotherapy program for upper [51, 52] and lower limbs [53, 70] in subjects with neurological or orthopedic disorders [54, 55] and for cognitive training [56].

2.1.2. Pre-processing and features extraction stage

The feature stage adopted in literature for human assessment aims to extract relevant features from the acquired motion trajectories. Nevertheless, a pre-processing filtering stage is mandatory to remove noise and outliers. This can happen when a subject moves out of the sensors range and/or Kinect data are inferred. There are cases when the Skeleton Tracking system does not have enough information in a captured frame to determine a specific joint position. In most cases, the system is still able to infer the joint position (resulting in "inferred state"). However, two types of noise are present in joint positions. One is the relatively small white noise caused by imprecision; the other is temporary spikes caused by inaccuracy, which happens when the joint has an inferred tracking state [71]. Some solutions can be adopted in this context to filter the motion signals from noise:

1. Smoothing filters: Auto-Regressive Moving Average (ARMA) filters, which represent a general class of filters. Specific smoothing filters for noise removal include Moving Average, Double Moving Average, Exponential Filter, Double Exponential Filter, and Savitzky-Golay filters.

2. Using joint tracking state in filtering: joints that are inferred are more likely to have temporary spike noises due to being less accurate. Also, the random noise levels are usually higher for inferred joints. The tracking state can be treated as the confidence level of the skeleton tracking system regarding the joint position [71]. Though the inferred joint positions are a very refined estimate of joint position, they may become inaccurate in some cases, depending on a person's pose. Therefore, one should expect that inferred joints have higher noise values, along with a possibility of a bias. The information of tracking state along with the motion tracking trajectories can be used to design an adaptive filtering stage.

3. Using the combination of more sources: the skeleton tracking algorithm does not take advantage of the sound source angle when the subject being tracked is speaking. In [72] and Extended Kalman Filter is designed to incorporate the information contained in the sound source angle for smooth out the jitter.

4. Using body kinematics to correct the filter output data: the anatomy and kinematics of a person provide valuable information that can be used to enhance the Skeleton Tracking joint accuracy. For example, some joints move or bend only in a certain direction, or some joints, such as hands, can move faster than other joints. This information represents useful constraints for bounding the motion data into the acceptable region.

2.1.3. Motion Assessment

The motion analysis in a telerehabilitation system, generally, is based on automated segmentation [27], identification [28] and assessment of movements employing machine learning or action similarity algorithms. In literature, human motion assessment approaches can be divided into two main categories [73]: *rule* and *template* based. In the *rule* based approach, experts (e.g., medical staff) identify some motion key descriptors, a set of rules (e.g., angles, joints position, relative distance, velocity), which define the "motion sample". The rule-based approach does not require the recording of exemplars and the dynamic building of ML models. In the *template* based approach, the sequence of gesture motion is *a priori* recorded and then used as an exemplar to be compared with the observations [73]. This comparison can be directly performed, via action similarity approaches (e.g., Dynamic Time Warping (DTW) [74, 55]), or using a machine learning model (e.g., Artificial Neural Networks (ANN) [75], Hidden Markov Model (HMM) [76]).

2.1.4. Rule-based methods

In [73], each exercise is described by rules to assess periodic movements and static poses. They define three types of rules for (i) dynamic movement, (ii) static poses and (iii) movement invariance. The rules are encoded using XML for its readability and extensibility. Afterwards, the dynamic rules are assessed in real-time by a finite state machine, while the static and invariance rules are evaluated comparing each frame supplied by the motion sensing device. In [77] and [78], the rules are defined in terms of the distance traversed of a set of joints for postural control and trunk flexion angle, and in terms of the trunk lean angle for gait retraining. Accordingly, in [79] the outcome measures of gait speed, step length and time, stride length and time and peak foot swing velocity were derived by Microsoft Kinect using supervised automated analysis. The validation of these features was performed with respect to a marker-based three-dimensional motion analysis. The method proposed in [80] integrates the Kinect with the inertial sensors by Kalman filtering. In particular, the knee angle and the ankle angle are used to assess the quality of sit-to-stand and squat, and the shoulder angle is used to assess the shoulder abduction/adduction quality. A rule-based intelligent control methodology is proposed in [81] to imitate the faculties of an experienced physiotherapist for a knee rehabilitation robot manipulator. The robot manipulator works based on impedance control and operates in two stages: teaching and therapy. If the patient resists against the motion the intelligent controller forces the knee to move up to the threshold limit defined in the database. In [82], the sit-to-stand exercise is evaluated by two metrics: (i) the minimum hip angle, in which a younger healthier person would typically have a larger value than an elderly, (ii) the smoothness of the head movement, which is computed as the area of the triangle that is determined by the second highest peak, the valley and lines that are parallel to the

axes on the head-speed-versus-time plot. In [83], the authors present an interactive physical therapy system for remote quantitative assessment of balance during the posture, measuring biomechanical features (i.e., joints angles and trunk sway) extracted by the Microsoft Kinect Skeletal Tracking system. The trunk sway measures and the joint alignment encapsulate the rules for motion assessment, while the minimum threshold is set empirically. More complex rules have been developed for the purpose of recognizing hand [84] and body gestures [85]. In [84], authors define a hand gesture by a sequence of monotonic hand segments. A monotonic segment refers to a sequence of hand configurations in which the angles of the finger joints are either non-increasing or non-decreasing. While in [85], a Gesture Description Language is proposed, in which a gesture is determined by a set of keyframe reporting the joint position of the Kinect skeletal tracking. Each rule is expressed in terms of one or more keyframes except the final rule, which defines the gesture in terms of a sequence of basic rules. A monitoring approach for home-based rehabilitation is proposed in [86] where a score is defined through the translation of medical requirements into quantitative rules, obtained by the physiotherapist performance. Accordingly, in [87], the assessment stage provides real-time scores related to the clinical targets of single rehabilitation exercises. The rules are designed via a fuzzy logic engine, according to the physiotherapist priority.

2.1.5. Template-based methods

Machine learning and time warping methods are widely used to identify [88, 89, 90, 91] and assess the human movement [74, 55, 92, 93, 94, 95, 96, 97].

Action similarity methods

DTW is used to align movements and to evaluate the action similarity between patient and exemplar (e.g. physiotherapist) [74, 55]. The DTW algorithm and fuzzy logic are employed in [55] to design a Kinect-based platform for rehabilitation exercises at home. The exercise, performed by the physiotherapist, is recorded as a reference sample for comparing the exercise performance of the patient. They use DTW algorithm to compare two sequences of different durations (i.e., motion trajectories of patient and physiotherapist) to determine the similarity between the standard and the patient exercises. In [74], authors propose an action tutor system which enables the user to interactively retrieve a learning exemplar of the target action movement and to immediately acquire motion instructions while learning it in front of the Microsoft Kinect. Basically, their system is composed of two stages: in the retrieval stage, non-linear time warping algorithms are designed to retrieve video segments similar to the query movement roughly performed by the user. Subsequently, in the learning stage, the user learns according to the selected video exemplar, and the motion assessment is performed by the joint difference and dynamic time warping algorithm. The action

similarity measurement proposed is based on cross-correlation, approximate string matching, DTW and a modified version of DTW that violates the boundary condition. In [92], a quantitative upper-limb evaluation for post-stroke rehabilitation is realized using a wearable inertial measurement unit. The proposed algorithm based on DTW is able to differentiate the level of limb function impairment, following Brunnstrom stage classification.

Machine learning methods

In [94], a non-invasive home monitoring system which extracts and analyzes the human motion and provides clinical feedback is presented. A quantitative evaluation of musculoskeletal disorders symptoms is performed, computing the detailed spatiotemporal objective measurements by a Temporal Alignment and a Spatial Summarization (TASS) in order to decouple the complex spatiotemporal information of multiple Skeletal Action Units (e.g., multiple repetitions of sit-to-stand movement) and to give a quantitative evaluation of musculoskeletal disorders. TASS is based on DTW and geodesic distance and the optimal alignment path indicates how two-time series match each other temporally. In [93], a graph-based method has been implemented to align two dynamic skeleton sequences acquired by Microsoft Kinect and to recognize tasks. In addition, an objective evaluation of the action performance has been realized by minimizing an energy function that jointly measures space and time domain differences. This measure has been used for recognizing actions, and for separating acceptable or unacceptable action performances. Authors of [95] analyze RGB images sequence to identify gait features, able to distinguish between normal and pathological patterns. In particular, they test different gait variables to compare performance results obtained by different machine learning approaches (i.e., K-nearest neighbors, Support Vector Machine (SVM) and Bayesian Classifier). Microsoft Kinect v2 Skeletal Tracking is used in [96] to evaluate motion impairments. The movement analysis is performed by comparing the test participants with an SVM model, generated from tracking movements of healthy people. Firstly, the algorithm recognizes motions, evaluating different machine learning methods (i.e., SVM, Random Forests, ANN, Gaussian Restricted Boltzmann Machines, Adaptive Boosting, LPBoost, RSUBoost, Total Boost, Bagging). Then, the analysis of mobility is implemented using multiple SVMs. SVM is also used in [97] to analyze the data acquired by a set of wearable accelerometers, in order to recognize different upper body motor tasks and predict the exercise intensity. In particular, the system uses a hierarchical algorithm, consisting of two layers of SVMs to first recognize the type of exercise being performed, followed by recognition of exercise intensity. The first layer uses a single SVM to recognize the type of the performed exercise. Based on the recognized type a corresponding intensity prediction SVM is selected on the second layer, specializing in intensity prediction for the recognized type of exercise.

2.1.6. Discussions

The main advantage of the *template* with respect to the *rule* based approach is the automatic assessment process that could be easily generalized to different types of exercise by acquiring reference sequences. On the other hand, the *rule* based method is less computationally expensive and provides a motion assessment with specific functional feedback (e.g., "Is the primary objective of the exercise reached?"), particularly useful in the rehabilitation context. In addition, the medical staff defines the rules of each movement according to the motor-functional scope and postural constraints of the exercise. These rules satisfy the invariance property of the movement and no normalization or scaling is needed. Nevertheless, drawbacks of the *rule* based approach are the lack of generalization and reusability for different exercises that lead to a large rule data set, requested for each motor task, difficult to synthesize within a telerehabilitation framework. However, the way to ensure a continuous monitoring of body motion and an accurate evaluation of rehabilitation therapy remains a challenge. Table 2.2 summarized the advantages and disadvantages of both methods.

Table 2.2.: Summary table of advantages and disadvantages of *rule* and *template* based methods

Features	*Rule based*	*Template based*
Computational effort	• It does not require the effort of model building and validation. However, when the complexity of the movement increases, it may be complex to precisely map the rule and to obtain an accurate movement assessment.	• The computational effort for ML methods mainly depend on the complexity of the model building (training stage). • The computational efficiency for action similarity techniques depends of the number of time series samples.
Data interpretation	• It can provide realtime feedback for both patient and clinician with much more salient and specific information regarding exactly how the motion deviates from the predefined gesture.	• Not all machine learning algorithm can provide interpretable score about the gesture correctness. • Action similarity technique can lead to a better outcome interpretation.

Features	*Rule based*	*Template based*
Calibration phase	• Since the rules can reflect the invariance of the gesture and it is independent from the subject who performs it no features scaling is needed.	• Features extraction stage is needed for compute features which are invariant among different subjects.
Automatism	• Physiotherapists and clinicians should carefully define the rules for each gesture expressing it in an implementable form. This would incur additional financial cost and effort, but prevent the definition of motion exemplar for defining the ground truth gesture. In addition it is also heavier the generalization of the set of rules for a different set of exercise.	• Physiotherapists and clinicians should define only the salient features. The ML model and/or the activity similarity techniques can be suitably generalized to different set of exercises.
Scalability	• The rules are often hard-coded into each application, making it hard to extend or modify an existing application. In the rehabilitation scenario, this drawback can be problematic because the exercises prescribed for different patients may have to be customized to meet the specific needs of each patient.	• The scalability and the customization of the algorithm for each different pathology need the design and the building of different ML models for each different patient condition. The model should be trained in different conditions (e.g., age, sex, pathological condition). • The activity similarity techniques need an additional mapping function to customize the obtained similarity measure across different conditions.

Features	*Rule based*	*Template based*
Robustness	• The influence of motion sensing error for the algorithm outcome depends on the definition and the tolerance of the set of rules against outliers.	• Adversarial machine learning [98] and/or robust machine learning method [99] should be taken into account to improve the performance of the movement assessment algorithm. • The outlier detection [100] methods should be employed during the pre-processing stage for the activity similarity approaches. In addition, also different similarity measures, more robust against outliers should be used.

2.1.7. Main Contribution

The novelty, introduced in the proposed approach, lies in combining aspects of the *rule* and *template* based methods, in order to overcome their drawbacks. According to a *rule* based approach, the proposed algorithm is able to provide quantitative scores related to kinematic features defined by clinicians. These features are evaluated by a Hidden Semi-Markov Model (HSMM) in order to perform the assessment stage. As for the template based method, the proposed approach can be easily generalized and reused for a different set of rehabilitation exercises, once the salient features of the motor task to be assessed have been selected. Note that the features are exercise-specific and need to be defined according to the exercise scopes. In particular, the physical exercises are modeled as a temporal sequence of postures, each depending only on the previous one so that they were formed by the continuous evolution of spatial configurations of the body posture [101]. This agrees with the Markov property and an HSMM can be exploited to model the physical mechanism and time constraints, related to each exercise.

2.2. State of art: Emotion Inference using Physiological predictors

The remainder of this section is organized as follows: the affective computing applications are shown in Section 2.2.1 with the emotional model adopted in literature and depicted in Section 2.2.2. The stimuli selection to elicit different emotions and the way to measure the affective state are described respectively in Section 2.2.3 and Section 2.2.4. Then, the emotion assessment task is presented in Section 2.2.5 focusing mainly with the analysis of physiological signals (see Section 2.2.6). In particular, more details about the affective signal processing are provided in Section 2.2.7. The major contributions in terms of machine learning method for solving the emotion recognition task are provided in Section 2.2.8, while author focus in the continuous recognition approaches in Section 2.2.9. The author discusses the challenge he aims to solve (see Section 2.2.11). Finally, the main contribution is described in Section 2.2.12.

2.2.1. Affective Computing applications

The ability of computers to understand, discern human emotions, and perform the appropriate actions is one of the key focus areas of research in Human-Computer Interaction (HCI). Hence, empowering computers and robots to understand human emotions would make HCI more meaningful and easier. Affective computing devices are being used in various domains such as education, healthcare, home automation, gaming, automotive and more. For instance, during online learning, the receptiveness of the student will be greatly increased if the computer knows the students' emotional state and provides the appropriate learning. A psychologist can diagnose the disease easily with the knowledge of the patient emotional state. Applications can be extended to missions involving very aged people, newborn, patients with autism etc., who will not be able to express their emotions explicitly. Picard describes three types of affective computing applications: 1) systems that detect emotions of the user, 2) systems that express what a human would perceive as an emotion (e.g., an avatar, robot), and 3) systems that actually "feel" an emotion. Detection, expression, and perception are crucial when designing technologies with affective capabilities [45].

Affective Computing in Healthcare

Previous research showed that in individuals affected by different pathologies (e.g., Asperger Syndrome (AS), High Functioning Autism (HFA)) the emotional responses are less differentiated, less positive and more negative than individuals without these disorders [102]. Individuals with AS or HFA experienced significant difficulties in the assessment and classification of their own emotions [103]. These behavioral traits can affect their relationships with other people [104]. Mobile applications like SymTrend

(SymTrend Inc., Cambridge, MA, USA) and Autism Track (HandHold Adaptive LLC, n.d.) have been developed, which allow patients with disabilities to enter behavioral data manually and to track changes over the time. These apps make more aware patients of their symptoms, give accurate advice and reminders. This allows patients and their therapist to have a picture of behavioral patterns, moods, and triggers that occur with any emotional outbursts. Post-traumatic stress disorder (PTSD), as defined by the American Psychiatric Association (APA), is the development of characteristic symptoms following exposure to actual or threatened death, serious injury, or sexual violence in one or more of the following ways: directly experiencing the traumatic event, repeated or extreme exposure to aversive details of the event/s, witnessing the event/s as it happens to another person; or learning the event occurred to a family member or close friend [105]. PTSD disables a person from carrying out daily activities and torments him/her with memories of stressful events [106]. Exposure therapy and stress inoculation are well-known techniques for treating individuals with PTSD [107]. Recently, researchers have included virtual reality in these therapies [108]. An example is StartleMart, a virtual reality-based gaming environment with stress detection capabilities that integrates cognitive behavioral approaches with physiological signals to treat veteran soldiers with PTSD [109]. This game simulates three highly stressful scenarios and the skin conductance is used to measure the body response to anxiety. A study conducted using StartleMart successfully correlated stressors on screen with peaks in skin conductance data, leading researchers to believe that these kinds of systems can help with the diagnosis and treatment of PTSD [109].

Affective Computing in Education

Nowadays, technology becomes a fundamental part of schools (e.g. smart boards, interactive presentations, etc.) in order to engage students and to provide a better education (teaching). Bringing affective computing into the classroom may enable students to get a personalized learning experience. However, have an individualized curriculum for each student is very time consuming and requires a depth understanding of what each student likes, dislikes, and his preferred learning method. Research shows that students who receive support from teachers and peers tend to feel more comfortable in school, enjoy school more, and participate more actively in classroom activities [110]. A visualization tool named EngageMe has been developed to support teachers in understanding how they are communicating with their students and how their pedagogical strategies can be improved to meet individual needs [111]. This system collects skin conductance data from students and graphs display the arousal level of each student to help the teacher reflect on his/her classes. To distinguish between the sources of these moments of arousal, a video feed is provided to the teacher, so he/she can determine if the arousals are due to classroom engagement or some other factor.

Affective Computing in Automotive

Affective computing can also be employed in the automotive area. Physiological signals are a useful metric for providing feedback about a driver state because they can be collected continuously and without interfering with the driver task performance. This information could then be used automatically by adaptive systems in various ways to help the driver better cope with stress. Some examples of this might include automatic management of noncritical in-vehicle information systems such as radios, cell phones, and onboard navigation aids. During high-stress situations, cell phone calls could be diverted to voice mail and navigation systems could be programmed to present the driver with only the most critical information or change to a quieter route to help reduce driver workload. In addition, the music selection agent might lower the volume or offer a greater selection of relaxing tunes to help the driver cope with their feelings of stress. Conversely, in low-stress situations, the car might recognize that more driver distractions could be tolerated and provide the driver with more entertainment options [112]. Moreover, Nass et al. [113] examine whether characteristics of a car voice can affect driver performance and affect. In an experimental study, participants had emotion induced by watching one of two sets of 5-minutes video clips. Participants then spent 20 minutes in a driving simulator where a voice in the car spoke 36 questions and comments in either an energetic or subdued voice. Participants interact with the car voice and when user emotion matched car voice emotion (happy/energetic and upset/subdued), drivers had fewer accidents, attended more to the road (actual and perceived), and spoke more to the car.

Recommender Systems

Inferring the affective state of the user during multimedia interaction is an important feature for producing, retrieving, transmitting, delivering and finally visualizing the right media material fulfilling the user expectations. Implicit tagging of videos using affective information can improve the performance of recommendation and retrieval system [114, 115, 116, 117, 118, 119] offering multimedia access not only based on semantic but also affective aspects of the interaction. Tkalčič et al. [120] presented a methodology for the implicit acquisition of affective labels for images. It is based on an emotion detection technique that takes as input the video sequences of the users' facial expressions. They extracted Gabor low-level features from the video frames and employed a kNN machine learning technique to generate affective labels in the valence-arousal-dominance space. They performed a comparative study of the performance of a Content-Based Recommender (CBR) system for images that uses three types of metadata to model the users and the items: (i) generic metadata, (ii) explicitly acquired affective labels and (iii) implicitly acquired affective labels with the proposed methodology. The results showed that the CBR performs best when explicit labels are used. However, implicitly acquired labels yield a significantly better performance of

the CBR than generic metadata while being an unobtrusive feedback tool bringing additional value to the former system.

Other Applications

Another field which can benefit from affective computing is Ambient Intelligence (AmI). AmI carries out a futuristic vision of living environments which are sensitive and responsive to the presence of people and, by taking care of their desires, intelligently respond to their actions improving their comfort and well-being. The Emotion-aware AmI (AmE) enhances the conventional idea of an intelligent environment by exploiting theories from psychology and social sciences for suitably analyzing the human emotional status and achieving a higher user satisfaction [121].

Other applications of affective computing include job interview performance, which is not only based on an individual knowledge but how well he can communicate that knowledge to the interviewer. Together with good verbal communication skills, the ability to control emotions is essential. MACH (My Automated Conversation coacH) developed at the MIT Media Lab, is a virtual agent that can read facial expressions as well as speech and language intonations. Verbal and non-verbal feedback is given to the users allowing them to improve their communication skills and control anxiety [122]. Moreover, due to the rise in security threats and controversies related to interrogation techniques, affective systems can unobtrusively detect specific emotions like anger, frustration, or deception in real time. Facial expressions recognition is currently as the primary measure to track emotional cues. Over the years, researchers have worked on developing an universal coding system for standard facial expressions, such as the Facial Action Coding System (FACS) [123]. Similarly, researchers developed the automated facial recognition system (AFERS) [124]. This system uses video streams and support vector machines (SVMs) to detect facial expressions, and the results can be viewed in real time. AFERS generates a graphical representation of the expressions over a period of time that helps investigators identify patterns of deception. Improvements of AFERS system also has the potential of being used in large gatherings like airports, games, or concerts to detect suspicious behavior in real time.

Innovations in computer game interfaces continue to enhance the experience of players. There have been technical advances that have driven game innovation over the past few decades, including advances in computer graphics, system performance, and human-computer interfaces. Recently, researchers have been interested in how the affective state of a game player can be brought into computer and video game experiences. Affective games, those that adapt or incorporate a player's emotional state tailoring the game responses, have shown promise in creating exciting and engaging user experiences. An example is the affective game engine developed by Negini et al. [125]. It adapts the player character and the non-player characters abilities and the environment via an affect-detecting middle-ware engine (AME) that translates

physiological inputs (GSR, HR) to game input. Their results suggested that adapting the game can increase player arousal and can automate balancing the difficulty of the game with the affective state of the player. In the study performed in [126] authors aim to understand engagement on the basis of the body movements of the player during a computer game. Preliminary results from two case-studies suggest that an increase in body movement imposed, or allowed, results in an increase in the player's engagement level. The relationship between body movement and game engagement is also investigated in [127].

2.2.2. Emotion model

In discrete emotion theory, all humans are thought to have an innate set of basic emotions that can be described as "discrete" because they are believed to be distinguishable by an individual's facial expression and biological processes [128]. The most famous discrete model was proposed by Paul Ekman and his colleagues in the cross-cultural study of 1992 [129]. They conclude that the six basic emotions are anger, disgust, fear, happiness, sadness, and surprise (see Figure 2.2 [129]).

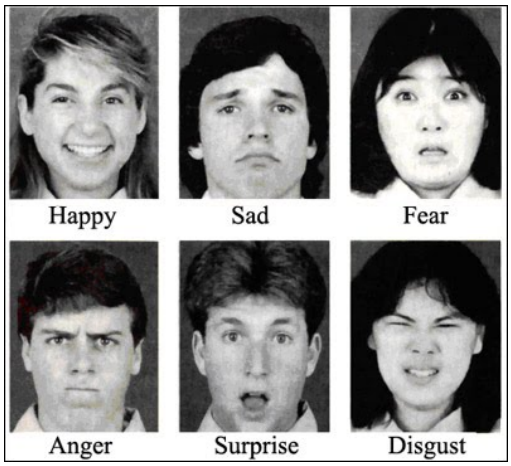


Figure 2.2.: Discrete model of Ekman

Ekman explains that there are particular characteristics attached to each of these emotions, allowing them to be varied and expressed in varying degrees. Each emotion acts as a discrete category rather than an individual emotional state (see also the Atlas of emotion [43]). Another discrete categorizations model of emotions proposed by Parrot is based on tree structure [130]. On the other hand, Russel summarized the cognitive structure of affect in a continuous model (see figure 2.3 [131]), where eight variables fall on a circle in a two-dimensional space in a manner analogous to points on a compass [131]. The horizontal (east-west) dimension in this spatial metaphor is the pleasure-displeasure dimension (i.e., valence dimension), and the vertical (north-south) dimension is arousal-sleep (i.e., arousal dimension).

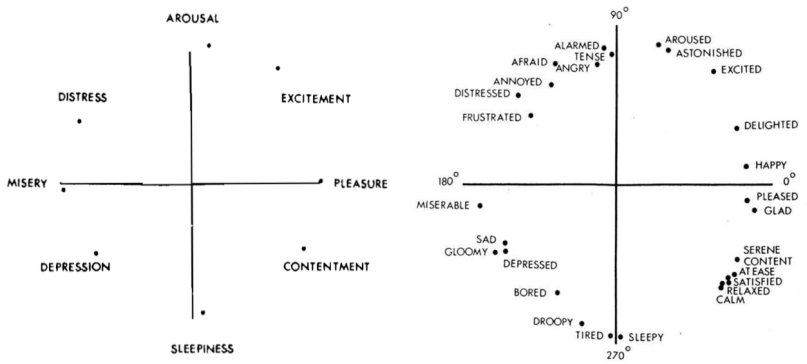


Figure 2.3.: Continuous model of Russell

The difference between the two orthogonal terms arousal and valence is pretty simple. Valence means the intrinsic attractiveness/goodness (i.e., positive valence) or averseness/badness (i.e., negative valence) of an event, object, or situation is positive or negative affectivity, whereas arousal measures how calming or exciting the information is. Russel's scale is widely used in research on affect, to quantitatively describe and assess emotions. While arousal and valence explain most of the variation in emotional states, the extension of the Russel model includes the dimension of the dominance [131]. Dominance ranges from a helpless and weak feeling (without control) to an empowered feeling (in control of everything). Plutchik [132] proposed a different hybrid model. In particular, he constructed a wheel-like diagram of emotions visualizing eight basic emotions: joy, trust, fear, surprise, sadness, disgust, anger, and anticipation. The wheel combines the ideas of circles representing emotions and a color wheel, where similar emotion (e.g. ecstasy and admiration) in the wheel are adjacent (see Figure 2.4 [132]). Emotions are also classified in this wheel according to a variety of intensities (e.g., ecstasy, joy and serenity).

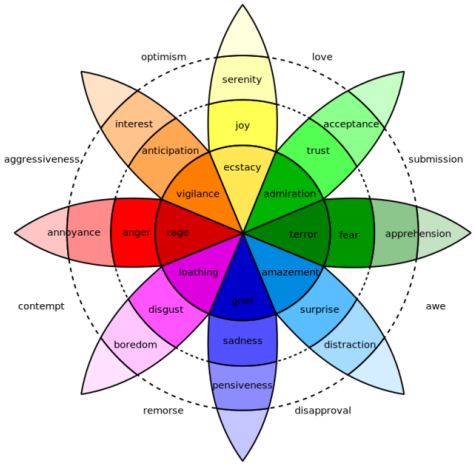


Figure 2.4.: Plutchik's wheel of emotions

2.2.3. Stimuli selection to elicit different emotions

In literature, several methods are used to elicit different realistic emotions. When the emotion is negative (i.e., negative valence), these methods should deal with ethic issues. Accordingly, the procedure for stimuli selection could lead to a strong computational effort [133]. Obviously, the elicited emotions have a very low intensity and the stimuli selection is performed in the laboratory, which is a completely different scenario with respect to real-life. The main stimuli used in this context are (please note that author chooses to show the main relevant database in the emotion recognition scenario):

- visual stimuli

  Pictures:

  - it is widely used the International Affective Picture System (IAPS) [134] with a dataset of more than 800 pictures.
  - the emotional database in [135]: they used the IAPS dataset and they acquired electroencephalogram (EEG), peripheral physiological signals, functional near infra-red spectroscopy (fNIRS) and facial video from 5 participants.

  Video:

  - Emotion elicitation using films [136]: they selected a subset of 16 films which successfully elicited amusement, anger, contentment. disgust, sadness, surprise, a relatively neutral state, and, to a lesser extent, fear.
  - MAHNOB HCI [137] database: it consists of two experiments. The response including, EEG, physiological signals, eye gaze, audio and facial expressions of 30 people were recorded. The first experiment was watching 20 emotional videos extracted from movies and online repositories while the second was tag agreement experiment in which images and a short video with human actions were shown to the participants first without a tag and then with a displayed tag. The tag was correct or incorrect and the participants' agreement with the displayed tag was assessed.
  - DEAP dataset [138] database: multimodal dataset with EEG and peripheral physiological signals of 32 participants recorded by 32 participants while watching a 40 one-minute long excerpts music videos. In this study, an extensive analysis of the participants' ratings is investigated.

- auditory stimuli:

  - Emotion recognition based on physiological changes in music listening [139]

  - International Affective Digitized Sounds (IADS) database [140]: collection of sounds able to elicit emotions.

- logical/mathematical operations and human-machine interaction:
  - Affective communication for implicit human-machine interaction [141]
  - Psychophysiological signals associated with affective states as related ti HCI [142]: Forty-three healthy students were exposed to computer-mediated stimuli, while wearable non-invasive sensors were applied in order to collect the physiological data. The stimuli were designed to elicit three distinct affective states: relaxation, engagement, and stress.
  - Emotion recognition from physiological signals using wireless sensors for presence technologies [143]
  - Emotion representation and physiology assignments in digital systems [144]
  - Bimodal emotion recognition using speech and physiological changes [145]

- combination of more methods (i.e., audio+pictures+human machine interaction)

Readers can found an extensive review of affective audiovisual database in [146, 147].

2.2.4. How to measure emotions?

The validity of self-reports of emotion is too often seen as an all-or-none phenomenon. Robinson and Clore [148] stated that the degree to which self-reports are valid varies by the type of self-report. Often, self-report of current emotional experiences are likely to be more consistent when the self-reports of emotion is made somewhat distant in time from the relevant experience [148]. An interesting review in [149] examined whether emotional states are associated with specific and invariant patterns of experience, physiology, and behavior. They suggested that measures of emotional responding appear to be structured along valence and/or arousal dimensions rather than discrete emotions (e.g., sadness, fear, anger). On the other hand, a physiological measure like GSR appears sensitive to arousal while facial EMG is sensitive to valence. In addition, these measures are not strongly related to one another. Practically speaking, then, there is no "gold standard" measure of emotional responding. For theories of emotion, this means that there is no "thing" that defines emotion, but rather that emotions are constituted by multiple, situationally and individually variable processes [149]. Mehrabian and Russell [150] proposed the Semantic Differential Scale for assessing the 3-dimensional structure of objects, events, and situations. It consists of a set of 18 bipolar adjective pairs that are each rated along a 9-point scale

(see [150] and [151]). However, one of the most used test for measuring emotion is the Self-Assessment Manikin (SAM) introduced in [151]. In particular, the SAM is a non-verbal pictorial assessment technique that directly measures the pleasure, arousal, and dominance associated with a person's affective reaction to a wide variety of stimuli (see Figure 2.5).

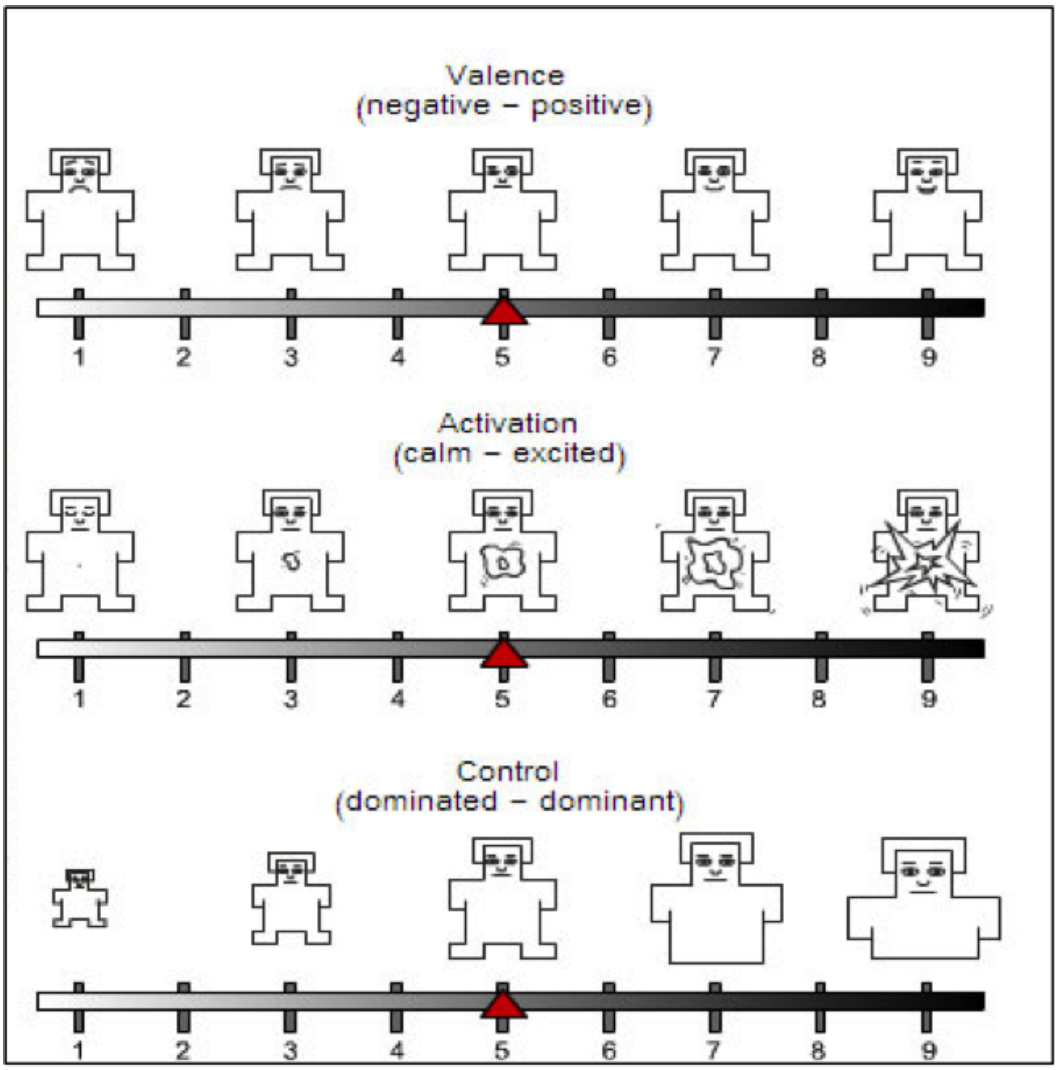


Figure 2.5.: The Self-Assessment Manikin (SAM) used to rate the affective dimensions in terms of valence (top panel), arousal (middle panel) and dominance (bottom panel)

The difference in judgments of dominance suggests that SAM might be more accurate in tracking the subject's feeling with respect to the Semantic Differential Scale [151]. Hence, several works [138, 152, 153] implemented the self-assessment method based on SAM to provide ground truth data (i.e., labels) for emotional states.

2.2.5. Emotion Assessment

Emotion assessment is often carried out through analysis of body expressions [154, 155, 156, 157, 158, 159] and/or physiological signals [160]. Other analyses include text [161, 162, 163, 164, 165, 166, 167], speech [168, 169, 170, 171, 172, 173] and body movements [174, 175, 176, 177]. Emotion recognition using facial expressions has many advantages and this reflects the effort of the researchers in this field. However, the facial expressions can be consciously controlled. For instance, a sad person can show a happy face to mask his emotions, which shows that facial expressions are not always linked with inner emotions. On the other hand, emotion recognition using text is applied to words or sentences in a particular language. Accordingly, it is widely known that people with different cultures talk with different tones. A person from a certain culture or area, talking in a normal tone might sound angry to someone from another culture. On the other hand, talking in a low and slow tone might be viewed as a sad emotion in one culture but polite and normal in another. This is a major drawback when it comes to developing a universal method for recognizing emotions using text and speech. However, to overcome this problem, researchers are moving towards the detection of emotions in multi-language text [167]. Emotion recognition through body movements and gestures is the least popular due to the difficulty to track the movement of the whole body with a simple camera. The unobtrusiveness of the device and the lower accuracy of the sensor should be taken into account in the real-world scenario. For this reason, the emotion assessment is carried out through analysis of peripheral responses without considering other signals such as electroencephalogram (EEG). Indeed the acquisition of EEG involves high cost and obtrusive equipment which influence the accuracy of data acquisition.

2.2.6. Physiological signals

The analysis of the physiological measurements is known to include emotional information as a response from the Central Nervous System (CNS) and the Peripheral Nervous System (PNS). Generally, physiological analysis has received less attention with respect to facial expression in the emotion assessment task, due to the complexity of the task and the obtrusiveness of the sensors used to acquire physiological measurements (i.e., electroencephalogram, electrocardiogram, galvanic skin response, skin temperature, electromyogram). However, since the physiological changes are controlled by the ANS providing information about the subjects' internal emotion and cannot be controlled consciously. The state of the art in this field includes affective signal processing and machine learning algorithm [160, 178, 179] for dimensionality reduction, features selection, and classification of different target emotions.

Although several studies focus the attention in recognize human emotion using physiological signals, the relation is not biunique. Table 2.3 resumes the outcomes of fifteen studies reported in [180] summarizing how physiological signals such as Gal-

vanic Skin Response (GSR), Heart Rate (HR), Skin Temperature (ST), Blood Pressure diastolic, Blood Pressure systolic and respiration are correlated with respect to the discrete emotion proposed by Ekman [129]. Numbers in parenthesis indicate how many studies support or oppose the named hypothesis (2-2-1 means 2 studies support the hypothesis, 2 do neither support nor oppose it, 1 oppose it).

Table 2.3.: Physiological features correlation:

	Fear	Anger	Sadness	Happiness	Disgust	Surprise
GSR	Increase (5-1-0)	Increase (1-1-0)	Decrease (1-0-0)	Decrease (1-0-0)	Increase (1-0-0)	n/a
HR	Increase (11-0-0)	Increase (8-0-2)	Increase (5-1-2)	Increase (3-1-1)	Increase (2-2-0)	Increase (1-0-0)
ST	Decrease (2-2-0)	Increase (2-1-0)	n.s. (2-0-0)	Increase (1-1-0)	Decrease (1-1-0)	n.s. (1-0-0)
BP diastolic	Increase (2-1-1)	Increase (9-0-1)	Increase (2-1-1)	Increase (4-1-1)	Increase (1-0-0)	n/a
BP systolic	Increase (4-0-0)	Increase (5-0-1)	Increase (3-0-1)	Increase (4-0-2)	Increase (1-0-0)	n/a
Respiration	Increase (3-0-0)	n/a	n.s. (1-0-0)	n/a	n/a	n/a

Dark green: strong evidence
Light green: some evidence
Yellow: no clear assumption can be made due to contradictory results or too few studies
Red: not sufficient evidence for either hypothesis (n/a - no studies available that provide sufficient evidence)

Emotion like fear discloses the strongest evidence that is correlated with respect to SC, HR, Blood pressure systolic and respiration. While surprise shows the lowest evidence that is correlated with SC, Blood pressure and respiration. Table 2.4 shows the physiological features correlation based with respect to the x-axis of the continuous model of Russell [131] (i.e., valence). The continuous scale was discretized in negative and positive valence. The table is based on the study performed in [180] on the reviews of 26 papers. Strong evidences were found that negative valence is correlated with respect to GSR, BP systolic and respiration. Table 2.5 shows the physiological features correlation with respect to arousal (i.e., the y-axis of the continuous model of Russell [131]) for the study reported in [180]. GSR, HR, BP systolic and Respiration shows the strongest evidence to be correlated with respect to the high arousal level.

Table 2.4.: Physiological features correlation (valence):

	Negative Valence	Positive Valence
GSR	Increase (8-2-0)	Decrease (1-0-0)
HR	Increase (26-3-4)	Increase (3-1-1)
ST	Decrease (7-4-0)	Increase (1-1-0)
BP diastolic	Increase (14-2-3)	Increase (4-1-1)
BP systolic	Increase (13-0-0)	Increase (4-0-2)
Respiration	Increase (4-0-0)	n/a

Dark green: strong evidence
Light green: some evidence
Yellow: no clear assumption can be made due to contradictory results or too few studies
Red: not sufficient evidence for either hypothesis
(n/a - no studies available that provide sufficient evidence)

Table 2.5.: Physiological features correlation (arousal):

	High arousal	Medium arousal	Low arousal
GSR	Increase (5-1-0)	Increase (2-1-1)	Decrease (1-0-0)
HR	Increase (11-0-0)	Increase (13-3-3)	Increase (5-1-2)
ST	Decrease (2-2-0)	Increase (3-3-1)	Not significant (2-0-0)
BP diastolic	Increase (2-1-1)	Increase (14-1-2)	Increase (2-1-1)
BP systolic	Increase (4-0-0)	Increase (10-0-3)	Increase (3-0-1)
Respiration	Increase (3-0-0)	n/a	Not significant (1-0-0)

Dark green: strong evidence
Light green: some evidence
Yellow: no clear assumption can be made due to contradictory results or too few studies
Red: not sufficient evidence for either hypothesis
(n/a - no studies available that provide sufficient evidence)

2.2.7. Affective Signal Processing

Obviously, data processing depends on several factors such as the sample rate of the signal and also the nature of the acquired signal. A deep review through the methods of signal processing and data analysis for emotion recognition is described in [179, 181], while other papers pointed out the most significant features used as predictors in the ML prediction algorithm [160, 182, 183, 184]. The affective processing stage can involve the following steps:

1. Pre-processing
    - Resampling: The idea behind resampling is the process of picking another sample in order to increase the sampling rate of the discrete signal to obtain a new discrete representation of the underlying continuous signal.
    - Artifact removal: Biosignals such as ECG, GSR and ST can also be inaccurate due to movement artifacts and also differences in bodily position. Techniques for detecting [100] and removing outliers [99] should be taken into account.
    - Filtering: filtering stage should be tailored to the specification of biomedical sensors and data acquisition procedure.
2. Synchronization & Segmentation: biosignals are synchronized and segmented based on events or stimuli.
3. Features extraction stage: features need to be extracted from the signals. The affective signals are processed in the time (e.g., statistical moments), frequency (e.g., Fourier), time-frequency (e.g. wavelets), or power domain (e.g., periodogram and autoregression). Often some features are computed also with respect to the baseline situation (e.g., a situation without stimuli).
4. Normalization: humans are known for their rich variety in all aspects. This difference lies also in the biosignals. Then, the developing of the classification stage required the normalization of the signals. Finding an appropriate normalization method is both important and difficult because it depends on factors that can easily change on a daily basis. Physiological signals can be normalized, baseline corrections (applied when comparing or generalizing multiple measurements from one individual across a variety of task), and range correction (reduce the inter-individual variance by a transformation that sets each signal value to a proportion of the intra-individual range). However, often the normalization of physiological features before learning the ML model is performed by the z-score approach.
5. Dimensionality reduction / Features selection: a crucial issue in machine learning problem for classification and regression is the identification of a representative set of features from which to design a classification/regression model for

a particular application. Dimensionality reduction techniques can be used to reduce the dimensionality of the features space. Supervised techniques utilize the class information in order to define the maximum separation criteria (e.g., Linear Discriminant Analysis), while unsupervised algorithm only considers the global structure of the data (e.g., Principal Component Analysis). Despite dimensionality reduction algorithms, features selection allow maintaining the interpretation of the features. Features Selection methods can be divided into two different categories: (i) filter and (ii) wrapper approaches. The former works independently on a classifier involved in pattern recognition, using different criteria to judge a feature set or to judge the performance of the classifier. While the latter consists in taking the estimated performance of a classifier as the proper feature selection criterion. Both approaches present advantages and drawbacks. Wrappers often reach better accuracy results of the learning algorithm, while filters are computationally less expensive and can be easily generalized without the need to re-implement the features selection stage when switching from one learning algorithm to another. In several applications, the advantages of filters overcome their disadvantages, and they can reach the same learning algorithm accuracy as wrappers.

6. Classification stage Supervised and unsupervised classification algorithm could be used to learn the emotional response, modeling the biosignal features. Machine learning methods include parametric and nonparametric techniques, as well as discriminative and generative models.

2.2.8. Works in literature

Table 2.6 summarized the most representative machine learning studies in the last 15 years employing peripheral physiological signals for recognizing different emotions state [185] and stress levels [186, 187, 188].

Table 2.6.: Works of literature: Emotion recognition using physiological signals, *n* is the number of subjects involved in the experiment

Ref	Year	Signals	n	FS/DR	Classifiers	Target	ACC %
[189]	2004	ECG, GSR, ST	50		SVM	3, 4 emotions	78, 62
[190]	2004	ECG, GSR, ST	50		NN	6 emotions	84
[191]	2007	ECG, GSR	40		Reg. Model	5 emotions	63-64
[192]	2008	ECG, GSR	72	Anova	SVM, NN	2 fun levels	70
[193]	2008	ECG, GSR	1	LDA	K-NN, SVM	4 emotions	67, 83
[194]	2009	ECG, GSR, ST	6		NN	4 emotions	88

Ref	Year	Signals	n	FS/DR	Classifiers	Target	ACC %
[195]	2010	ECG, GSR, ST	34		K-NN, NN	3 emotions	65, 83
[183]	2011	ECG, GSR	19	SBS	LDA	2 boredom state	94
[196, 197]	2011	ECG, GSR	80		Fuzzy	2 stress levels	99
[198]	2011	ECG	42		LDA	2 stress levels	90
[199]	2011	GSR	9		SVM	2 stress levels	78, 73
[200]	2012	ECG	44		LDA	2 valence ratings	89
[138]	2012	ECG, GSR, ST RESP, EMG	32	Fisher	NB	2 valence, arousal ratings	57, 63
[201]	2013	ECG	40		k-NN	2 stress levels	94
[202]	2013	GSR, Accel, PHO	18		SVMs	2 stress levels	75
[203]	2014	ECG	39		HMM	5 stress levels	96
[204]	2014	HR, GSR, OXY	101	LSD	RF	5 emotions	74
[205]	2014	HRV	8		SVM	2 mood states	96
[206]	2015	HRV	27	Friedman	QDA	2 valence, arousal ratings	85, 84
[207]	2016	ECG	21, 5		RF	2 stress levels	92
[208]	2016	HRV	22		Fuzzy ARTMAP	3 stress levels	75, 80
[209]	2017	RR, GSR, ST	10		K-NN	2 stress levels	84
[209]	2017	RR, GSR, ST	10		DT	4 emotions	90

2.2.9. Continuous recognition approaches

The continuous recognition approaches require to label data continuously in real time by different annotators. The annotators could use a sliding controller to annotate both emotional dimensions separately. In order to solve the continuous recognition task, several sequential learning approaches have been applied in literature [210, 211, 212]. The Long Short-Term Memory Recurrent Neural Networks was applied in [210] for adequate prediction of emotion in a three-dimensional space modeling long-range dependencies and capturing emotional history. Accordingly, the temporal evolution of emotion is also evaluated for discrete labels employing Conditional Random Fields. The approach presented in [211] fuses facial expression, shoulder gesture, and audio cues for dimensional and continuous prediction of emotions in valence and arousal space. The bidirectional Long Short-Term Memory neural networks (BLSTM-NNs) and Support Vector Machine regression were applied to solve this task. The BLSTM-NNs approach outperforms SVR due to their ability to learn past and future emotion response. An Output-Associative Relevance Vector Machine regression framework that augments the traditional RVM regression learning non-linear input and output

dependencies is developed in [212]. Their methodology is consistent to model temporal and spatial dependencies. Dimensional models are considered important in the emotion recognition task as a single label may not reflect the complexity of the affective state conveyed by a facial expression, body gesture or posture [211]. However, the continuous recognition approaches may not be always suitable in the real-life scenario due to lack of difficulty in collection continuous labels

2.2.10. Multiple Instance Learning

Current MIL methods have proved to be useful in a variety of domains, ranging from bioinformatics [213], medical image analysis [214], text processing [215], educational scenarios [216] and object recognition and tracking studies [217]. In the affective image analysis, the MIL was proposed for modeling the spatial ambiguity of the emotion [218]. They firstly extracted blocks of an image at multiple scales using different image segmentation methods and represent each block using the bag-of-visual-words method. Then, MIL was used to predict the dominant emotion of the image. The authors in [219] applied MIL to the problem of automatic pain recognition from video. They represented each video as a bag containing multiple time-segments which are modeled using MIL. Their work encapsulates the temporal dynamics by representing the data not as individual frames but as segments. Accordingly, in [220] the MIL was used for the music emotion recognition in order to automatically recognizing the affective content of a piece of music. They captured the music emotion dynamics using a multi-instance structure based on song-segment-sentence. Recently, MIL was employed for behavioral coding [221] during problem-solving discussions. In particular, they treated each discussion as a collection of short-term behavioral expressions which are manifested in the acoustic, lexical, and visual channels. Their framework allows revealing the local/global nature of behaviors, estimating the level of ambiguity presented via a particular channel.

Compared to the works reported above, the MIL based application aims to predict the valence/arousal state modeling both the dynamic and the ambiguity of the emotion over time. Additionally, differently from the continuous emotion recognition approaches ([210, 211, 212] and reference therein), it provides a solution for learning in presence of a weakly supervised setting. This is often the case of real-life applications where the sparsity of the label and the lower accuracy of the signal should be considered.

MIL background

In the MIL paradigm, the learner receives a set of *bags* along with the corresponding label. Each bag contains multiple instances. Within this paradigm, the data is assumed to have some ambiguity in how the labels are assigned. A bag is labeled negative if all of its instances are negative, while a bag is labeled positive if there is at least one positive instance (see Fig. 2.6).

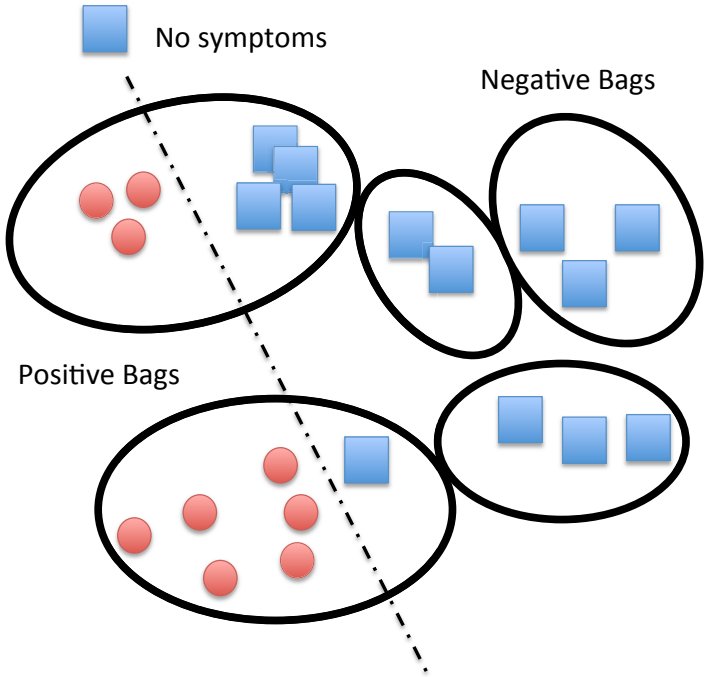


Figure 2.6.: The intuitive idea behind MIL

The MIL task is to predict the label of bags or the label of instances (see Fig. 2.7).

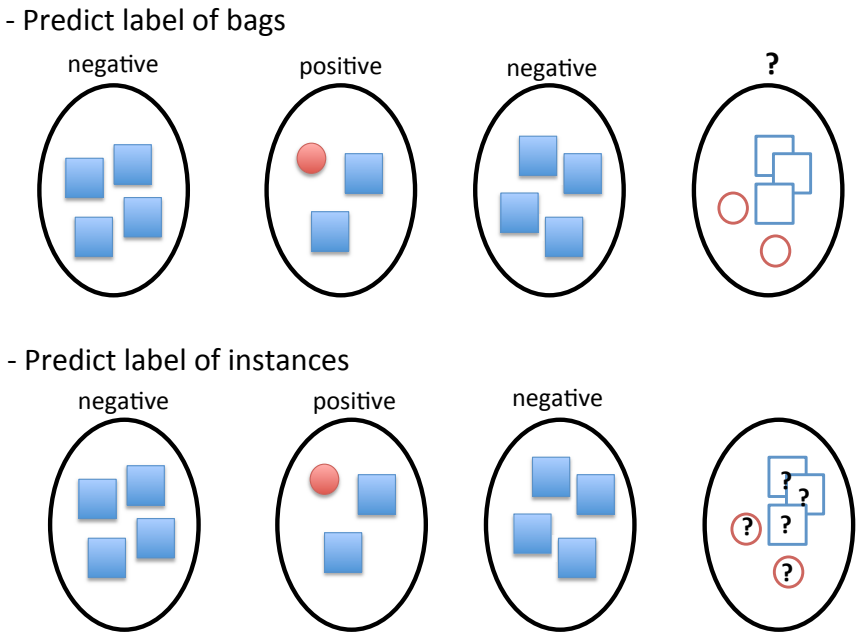


Figure 2.7.: The MIL task

The MIL problem was originally formalized for *drug activity prediction* by [213], in which authors developed a multiple instance algorithm for learning Axis-Parallel Rectangles (APRs). Afterwards, in [222] authors introduced the Diverse Density

framework for solving person identification problem. The combination of EM with the Diverse Density algorithm [223] allowed improving the computation time and the robustness against the number of irrelevant features. Accordingly, also several adaptations of Support Vector Machine (SVM) were proposed for MIL. In the Normalized Set Kernel configuration of [224] a traditional SVM is trained with bags represented as the sum of all its instances, normalized by its 1 or 2-norm, while in the Statistical Kernel set [224] every bag is transformed into a feature vector representation (i.e., the maximum and minimum value across all instances in the bag). Others SVM approaches called respectively maximum pattern margin and maximum bag margin, follow the heuristic approximation performed by [215]. The maximum pattern margin formulation (mi-SVM) aims to relabel the instances in positive bags using the learned decision hyperplane. In particular, if a positive bag contains no instances labeled as positive, the instance that gives the maximum value of the decision function for that bag is relabeled as positive. While in the maximum bag margin formulation (MI-SVM), for every positive bag, the learned decision function is used to select the bag instance that gives the maximum value. A different approach based on balancing and transductive constraints was presented in [225]. Unlike other SVM-based MIL methods, their algorithm is particularly effective when the positive bags are sparse (i.e., contain few positive instances). In the following subsections, the application of MIL for the emotion recognition task is proposed during a multimedia interaction, using the physiological features.

2.2.11. Discussions

The problem of emotion recognition from physiological signals during multimedia interaction has been the subject of several papers, see [138, 152, 226, 139, 227, 228, 229, 230] and references therein. User factors have been partially neglected, due to the inherent difficulty in dealing with individual differences and in measuring (unobtrusively) individual characteristics. In the context of real usage, the modern, low cost and unobtrusive wearable technologies (i.e., smartwatch) open a new realm of possibilities to estimate the emotional state in different scenarios. During recent years, several databases of physiological measurements in affective computing tasks have been collected and released [231, 138, 232], providing high-quality data for learning and benchmarking state inference models. In this context, very good results were obtained in highly controlled experiment setups where the stimuli evoke strong emotional responses [139, 160, 178, 153]. Accordingly, in the less controlled setups, the ground truth emotion labels come directly from user evaluations and not always the computed results are comparable with those obtained in [138, 152, 183, 233, 234, 229, 230, 189, 227, 235, 228]. Traditional approaches [138, 139, 229, 230, 236] have considered each video as a single instance and have employed ML model based on video-level features, assuming that the affective state

is consistent over the entire video [237]. However, since the emotion elicited by multimedia interaction is time-varying [238, 239] this assumption may be violated. Accordingly, not all observation windows have the same predictive power: it is possible to identify time interval which leads to better prediction of the presence of the judged emotion. This means that a specific physiological response can occur locally. In other words, if the user reports a positive/negative valence for the entire video, it does not imply that all the multimodal stimuli are perceived as positive or negative-specific physiological responses can occur locally. This ambiguity of the label reflects how the ground truth cannot always be continuous and reliable, reflective a summative global response. Especially in the real-life application the continuous labeling of emotion is critical: the labeling of the data is sparse and possibly describing only the more important emotional events.

2.2.12. Main Contribution

The main contribution of the proposed approach is the introduction and the application of a framework from the machine learning literature called Multiple Instance Learning (MIL) [213]. This choice offers a viable and natural solution for learning in a weakly supervised setting, taking into account the variability and the ambiguity of emotional response. The work contributes to the *affective computing field* as follows:

- it introduces the MIL framework for capturing the dynamic nature of emotion [240, 241]. The current approach seeks to discover the most prominent emotional events rather than the continuous affective changes that occur.

- it proposes an application of three MIL-based methods for the emotion recognition task and demonstrates significantly improved prediction performance over the best standard supervised machine learning approaches.

- it measures and demonstrates the reliability of the proposed approach towards the real world usage, where the unobtrusiveness of the device and the lower accuracy of the sensor should be taken into account.

The key findings, as well as the presented methodology, is general: it can be applied for a different set of observable cues (e.g. physiological signals, facial expressions, body gestures and movement, speech). However, in this work, we chose to focus on physiological signal since they cannot be easily faked or suppressed, and can provide direct information about the user's affective state [240].

2.3. The emotional effects on movement execution

In recent years there was an increasing demand to develop telerehabilitation system in non-clinical settings such as the home or workplace. These systems offer an opportunity for an individualized rehabilitation program and is based on regular monitoring of the patient's progresses respect to the treatment aim and subject's expectation [242, 243]. However, in such systems, the possible emotional effects on movement execution have been largely neglected. In the rehabilitation scenario, the pain-related emotions can influence the human movement and they represent a major barrier to effective self-rehabilitation in chronic pain [244]. Chronic pain is defined as pain that persists despite the resolution of injury or pathology or with no identified lesion or pathology [245]. The emotional response could generate anxiety that can cause marked reluctance to undertake therapies which are perceived as potentially exacerbating pain to the extent of avoiding them [244, 246, 247]. In [248] authors developed a Support Vector Machine framework combining as predictors a fusion of body motion and muscle activity descriptors to discriminate three levels of pain. The salient features were identified by a backward features selection approach. An understanding of how chronic pain and chronic pain related emotions are expressed is investigated in [244]. In addition, they provided a multimodal fully labeled dataset (EmoPain) for the chronic lower back pain. The dataset contains naturalistic pain-related affective expressions (facial and vocal) and behaviors (movement and muscle activity) of people suffering from chronic lower back pain while carrying out physical activity. Finally, they provided preliminary experimentations to investigate the possibility of automatically recognizing the facial expression of pain and pain-related body movement. The former is identified solving a classification task by Support Vector Machine, while the latter is analyzed building a regression model through the random decision forest. Their performance evaluation has been carried out using the leave one subject out procedure.

Chapter 3.

Materials

3.1. Quantitative Assessment of Human Movement

The validation of the adopted sensor is provided in Section 3.1.1. While the population enrolled as well as the description of the collected dataset is reported respectively in Section 3.1.2 and Section 3.1.3.

3.1.1. Validation of the adopted sensor

In the literature, the accuracy of Kinect v1 sensor is extensively investigated whereas the accuracy of Kinect v2 sensor is currently of utmost concern. The accuracy of Microsoft Kinect v1 was analyzed in the context of movement analysis with respect to movement artefact [249] or to gold standard systems (i.e. stereo) during different motor tasks such as gait analysis [250, 251, 252, 78], static [253, 254, 255] and dynamic postures [256, 257, 50, 258, 259].

The rehabilitation exercise monitoring involves considering dynamic movements with a wide range of motion and issues related to the joints tracking such as joints occlusion due to postures adverse to the vision sensor. For this reason, the accuracy analysis of Kinect v2 is provided in terms of joint positions and angles during dynamic postures used in low-back pain rehabilitation [260]. In particular, the joint positions and angles represent clinical features, chosen by medical staff, used to evaluate the subject's movements in a telerehabilitation scenario. These features represent the motion descriptor adopted by clinicians to assess the performance of patients during the execution of each exercise. In this context, several vision-based approaches for telerehabilitation employ a features extraction stage in order to extract clinical features, which are able to quantify the quality of different physical exercises execution [86, 261, 262, 87]. The accuracy of Kinect v2 is investigated with respect to the gold standard represented by a stereophotogrammetric system characterized by 6 infrared cameras (i.e. Elite motion capture system BTSEngineering). The results provide salient information for evaluating the reliability of Kinect v2 sensor for dynamic postures.

Population

Twelve healthy non-athletes young subjects ($mean \pm std$: 27.6 ± 2.5) who did not report neurological or musculoskeletal problems, no recent trauma, and who did not perform competitive sports were recruited for the experiment. The study protocol was conformed to the Helsinki protocol for clinical trials and was approved by the local ethics committee. All subjects signed the informed consent before taking part in the study.

Exercise description and features extraction

The exercises are selected by clinicians and they are widely used for low back pain physiotherapy. Two exercises involve the upper body and are respectively the lifting of arms (see Fig. 3.1.a), and the lateral tilt of the trunk (see Fig. 3.1.b). An exercise involves lower body: the squatting (see Fig. 3.1.c). Subjects were asked to repeat each exercise 6 times consecutively. Each exercise was performed with the subject in front of Kinect v2, starting by the upright position. Before starting the exercises, each subject has to perform a specific movement (i.e. lifting both the arms) in order to synchronize Kinect v2 with the stereophotogrammetric system.

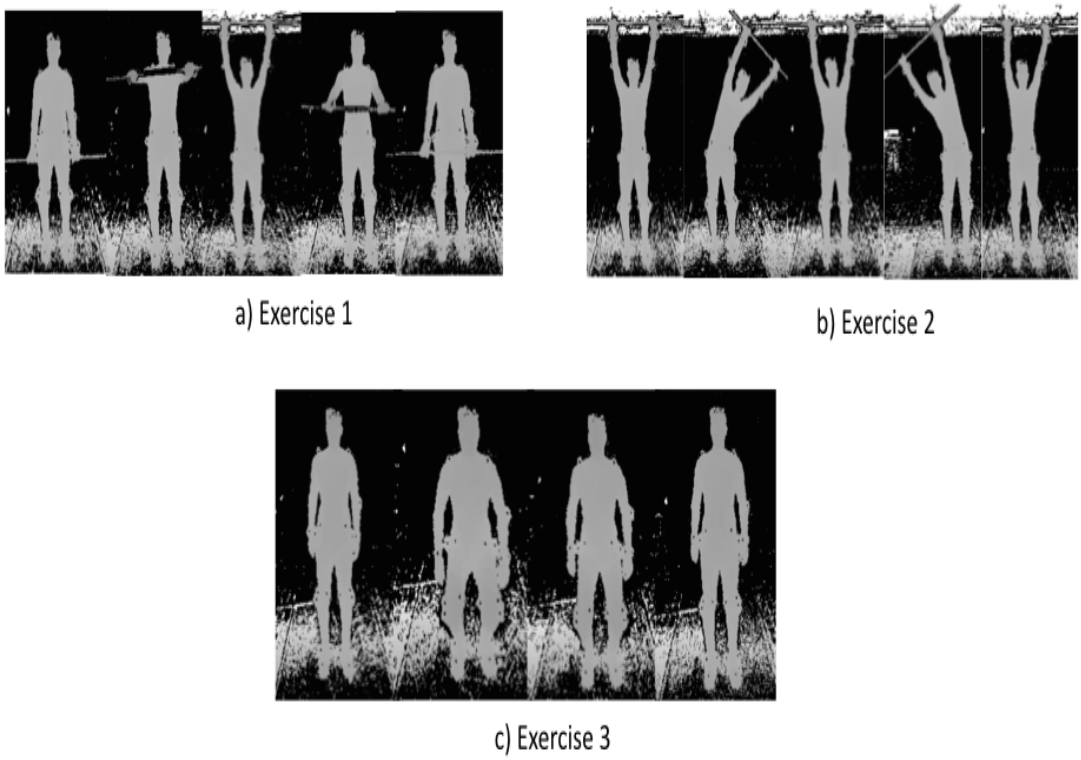


Figure 3.1.: Physical exercises widely used for low back pain physiotherapy involving upper body (a-b) and lower body (c)

The Kinect v2 accuracy is measured through salient features considered clinically

relevant descriptors of subject's performance. For each exercise, these clinical features are classified as dynamic features (DF) and static features (SF) (see Fig.3.2). DF describe the kinematic goals that subjects have to reach (i.e. lifting of arms or arm lateral tilting), while the SF represent the multi-joint posture the subject has to maintain during the execution.

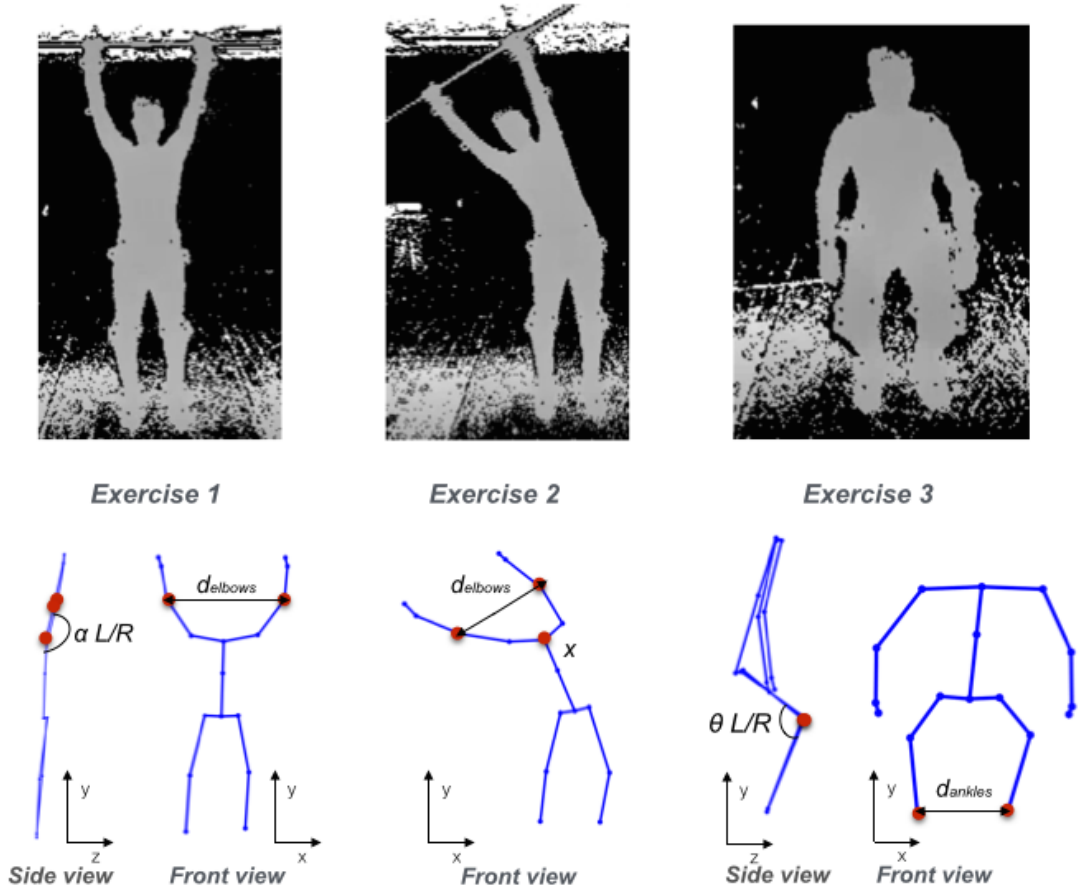
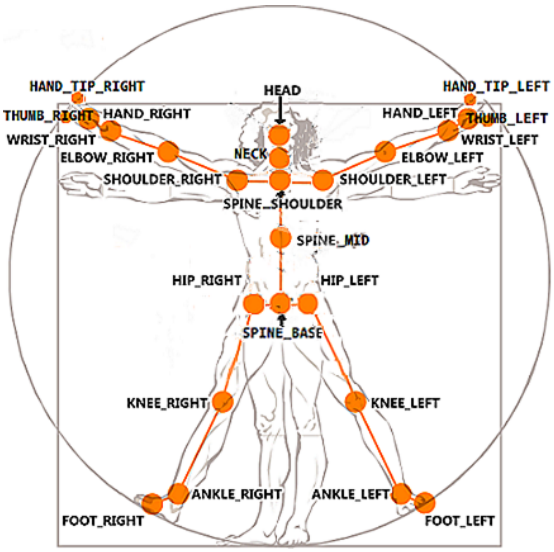
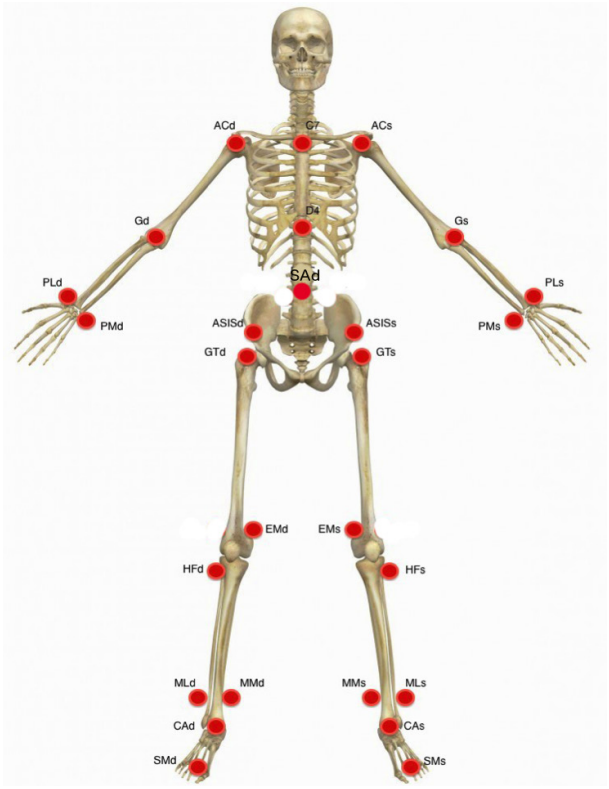


Figure 3.2.: Clinical features extracted for the three exercises

The kinematic protocol, adopted for the stereophotogrammetric system, is shown in Figure 3.3.b and was defined to be closely related to the Kinect joints skeleton (i.e., Fig. 3.3.a).



(a)



(b)

Figure 3.3.: The 25 joint positions provided by Kinect v2 (a). Marker locations of the stereophotogrammetric system (b).

Table 3.1 shows the extracted features for each exercise, computed with the relative landmark joints, measured with Kinect v2 and stereophotogrammetric system. The torso oscillation x, computed in Exercise 2, is normalized to zero mean. Maximum

and minimum of x represent the oscillation in right and left side, respectively.

Table 3.1.: Clinical features evaluated

Exercises	Features type	Clinical Features	Kinect v2 Skeleton Joint	Elite Marker
1) Lifting of the arms	DF	Underarm Angles ($\alpha_{R/L}$)	Spine Shoulder Elbow Right Elbow Left	C7 Gd Gs
	SF	Euclidean Distance between Elbows (d_{elbows})	Elbow Right Elbow Left	Gd Gs
2) Lateral tilt of the trunk	DF	Torso Oscillation (x)	Spine Shoulder	C7
	SF	Euclidean Distance between Elbows (d_{elbows})	Elbow Right Elbow Left	Gd Gs
3) Squatting	DF	Knee Angles ($\theta_{R/L}$)	Hip Right Knee Right Hip Left Knee Left	ASISd mean(EMd,HFd) ASISs mean(EMs,HFs)
	SF	Euclidean Distance between Ankles (d_{elbows})	Ankle Right Ankle Left	mean(MLd,MMd) mean(MLs,MMs)

Data processing and analysis

The 25 skeleton joint positions of Kinect v2 are recorded at 30 fps. The counterpart joint locations identified by the stereophotogrammetric system are recorded at 50 fps. Calibration error of the motion capture system was calculated before each experimental session. Among sessions, the mean error of the stereophotogrammetric system was 1.57 ± 1.1 mm. A cubic spline-interpolation is implemented for solving occluded marker problem. Both stereophotogrammetric and Kinect v2 data are filtered with a low-pass Butterworth filter with cut-off frequency $f = 5$ Hz in order to filter temporary spikes as artifacts and noise. The stereophotogrammetric data are downsampled at 30 fps. Spatial accuracy is defined by the differences between the features calculated with Kinect v2 and the stereophotogrammetric system respectively. For the DF the absolute error is computed for each maximum and minimum peak, while for the SF the difference is computed in terms of offset and Root Mean Square Error (RMSE) after removing the spatial offset. Moreover, for each exercise, the temporal accuracy is evaluated comparing the time-difference between each maximum-peak of DF.

Sensor Validation Results

Accuracy results are provided comparing the clinical features, extracted by Kinect v2, and the same features obtained by the stereophotogrammetric system considered as the reference value. Then, the spatial and temporal accuracy of Kinect v2 is investigated.

Spatial Accuracy

Figure 3.4 shows the distance comparison between the DF of Exercise 1 (α_R) calculated for both the systems together with the peak Absolute Error (AE) and peak Relative Error (RE). Features computed by Kinect v2 follow the trend of the ground truth signal; in particular, Kinect v2 overestimates the maximum peak while for the minimum the two values are comparable. Figure 3.5 shows the comparison between the SF of Exercise 1 (d_{elbows}). Although there is a systematic offset, removed in the figure below, it can be noticed a similar trend of this feature for both the systems.

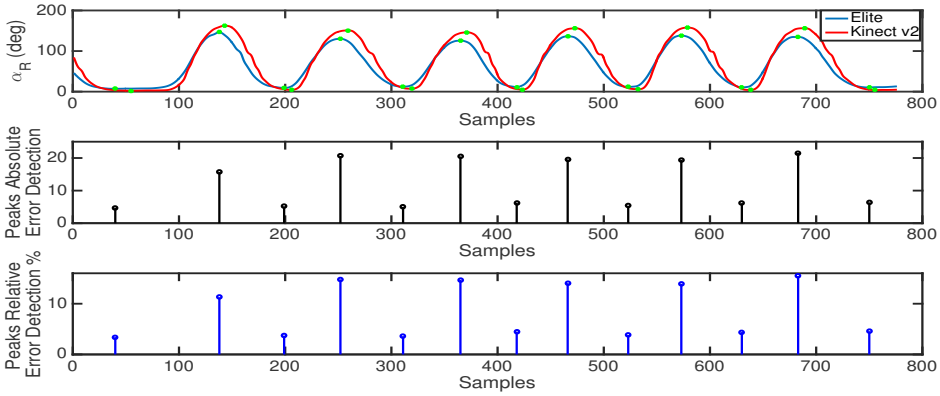


Figure 3.4.: Comparison between right underarm angle (DF Exercise 1: α_R) computed by Kinect v2 and stereophotogrammetric system

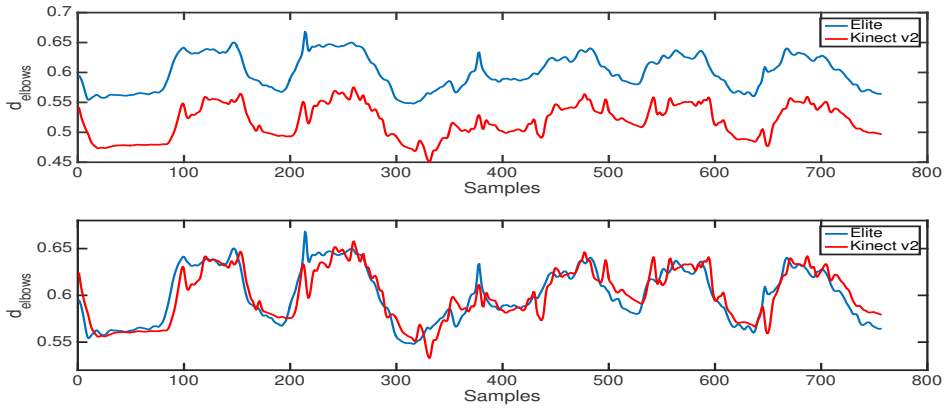


Figure 3.5.: Comparison between elbows distance (SF Exercise 1: d_{elbows}) computed by Kinect v2 and stereophotogrammetric system

Table 3.2 shows the spatial accuracy for each DF in terms of peak AE and RE In the Exercise 1 the maximum error is measured for the maximum right underarm angle ($Max(\alpha_R)$) during the lifting of the arms: RE=12.1%, AE=18$\pm$9.5. For Exercise 2 the error during the maximum ($Max(x)$, tilting in the right side) and minimum ($Min(x)$, tilting in the left side) oscillation is comparable: respectively RE=12%, AE = 5.7$\pm$2.2 and RE=12.7%, AE=6$\pm$2.5. The maximum error for Exercise 3 is measured during knee flexion for both legs ($Min(\theta_R)$, $Min(\theta_L)$): respectively RE=24.3%, AE=24$\pm$ 10.4 and RE=26.3%, AE=26$\pm$8.1.

Table 3.2.: Spatial accuracy comparison of the DF

Exercises	DF	AE ($\mu \pm \sigma$)	RE (%)
1) Lifting of the arms		(deg)	
	$Max(\alpha_R)$	18$\pm$9.5	12.1
	$Max(\alpha_L)$	13.1$\pm$9.4	8.7
	$Min(\alpha_R)$	11.4$\pm$6.4	7.5
	$Min(\alpha_L)$	8.1$\pm$5.3	5.3
2) Lateral tilt of the trunk		(cm)	
	$Max(x)$	5.7$\pm$2.2	12
	$Min(x)$	6$\pm$2.5	12.7
3) Squatting		(deg)	
	$Max(\theta_R)$	9.5$\pm$5.9	9.7
	$Max(\theta_L)$	8.2$\pm$2.7	8.4
	$Min(\theta_R)$	24$\pm$10.4	24.3
	$Min(\theta_L)$	26$\pm$8.1	26.3

Table 3.3 shows the spatial accuracy computed for each SF in terms of offset and RMSE. In the Exercise 1 and 2 the offset associated with the distance between elbows (d_{elbows}) is about the $\approx$ 8 cm, while the RMSE is respectively 2.7 cm for the Exercise 1 and 4.7 cm for the Exercise 2. In the Exercise 3, the offset, related to the distance between ankles, is 3 cm, while the RMSE is 2.4 cm.

Table 3.3.: Spatial accuracy comparison of the SF

Exercises	SF	Offset	RMSE
1) Lifting of the arms		(cm)	(cm)
	d_{elbows}	8.3	2.7
2) Lateral tilt of the trunk		(cm)	(cm)
	d_{elbows}	7,8	4.7
3) Squatting		(cm)	(cm)
	d_{ankles}	3	2.4

Temporal accuracy

Temporal accuracy is evaluated in terms of number of frame difference. Then the temporal distance between two maximum local peaks of DF is considered. Table 3.4 shows the time-peak distance between the two systems. In this case a positive number of frame difference indicates an overestimation of the Kinect measure related to the time between repetitions (two consecutive peaks).

Table 3.4.: Temporal accuracy comparison of the DF

Exercises	DF	AE ($\mu \pm \sigma$)
1) Lifting of the arms		(# $frames$)
	time-peak distance (right)	1.4 ± 3.4
	time-peak distance (left)	1 ± 3.4
2) Lateral tilt of the trunk		(# $frames$)
	time-peak distance	1 ± 5.5
3) Squatting		(# $frames$)
	time-peak distance (right)	0.5 ± 1.1
	time-peak distance (left)	0.5 ± 2.1

Results of temporal accuracy show as Kinect v2 could accurately measure timing characteristic of physical exercises confirming the results in [250, 254]. Regarding the spatial accuracy, the Kinect v2 sensor is able to reproduce salient D.F. in a comparable manner with those obtained from the stereophotogrammetric system. The novel Microsoft Kinect sensor seems to ensure a better accuracy in the motor task involving upper limbs (i.e., Exercise 1: maximum R.E.=12.1% and Exercise 2: maximum R.E.=12.7%) with respect to task involving lower body (i.e., Exercise 3: maximum R.E.=26.3%). Therefore, a relevant systematic error appears in the Exercise 3 during the minimum knee angle (maximum knee flexion), while the Exercise 1 and 2 show more error variability. This is probably due to wider translational and angular excursion during the Exercise 1 and 2, with a larger measurement volume respect to that involved in the Exercise 3. In particular, during the maximum knee flexion in the Exercise 3 some salient joints used to compute the clinical features seem to be estimated by Kinect v2 with a systematic bias respect to the gold standard system. A possible systematic error could be not very significant for the continuous and overall temporal evaluation of the patient. Moreover, results in Exercise 2 suggest how the magnitude error is comparable during right and left oscillation (i.e., R.E.=12% and R.E.=12.7%, respectively) highlighting a vertical symmetry. Accuracy-related to S.F. discloses as Kinect v2 follows the trend of the related features, measured by the stereophotogrammetric system. In particular, the offset related to the distance between elbows extracted in Exercise 1 is maintained also during Exercise 2 (i.e., ≈ 8 cm), where the elbows

have to be held in the same position. While for Exercise 3 the offset related to the distance between ankles is smaller (i.e., 3 cm).

3.1.2. Population

The experiment was conducted at the Neurorehabilitation Clinic of the University Hospital of Ancona (Italy), where subjects were recruited to perform selected exercises in front of the Microsoft Kinect v2 sensor. From May 1st through 30th 2016, clinicians enrolled, in the study, 22 Healthy Subjects (HS), free from disabling musculoskeletal or neurological problems, back pain, and recent trauma. Moreover, 19 patients suffering from chronic disabilities due to Neurological or musculoskeletal Disorders (ND) were enrolled among those consecutively referred to the Neurorehabilitation facility. They matched the following inclusion criteria: subjects suffering from low back pain and axial (trunk or gait and balance) disabilities, without severe dementia, walking with some limitations or under supervision (Walking Handicap Scale Category ≥ 3 (=limited household walker) ≤ 5 (=Least-limited community walker)), and not suffering from an acute phase of pain or recent trauma. The flow of the enrolled subjects through the study is reported in Figure 3.6. The study was carried out in conformity with the Helsinki protocol [263] for clinical trials and was approved by the local ethics committee of the University Hospital. All subjects performed the experimental exercise protocol after signing the informed consent.

3.1.3. Exercises

The clinicians selected the following five exercises as a case study [264]. Exercises #1-#4 involve the upper body segments, and include: lifting the arms (Figure 3.7a), tilting the trunk sideways with extended arms (Figure 3.7b), rotating the trunk (Figure 3.7c), rotating the pelvis rotations in the transverse plane (Figure 3.7d). The Exercise #5, the squat (Figure 3.7e), mainly involves lower body segments. The exercise selection was made for both clinical and technical reasons. Firstly, they are very basic motor tasks aimed at improving axial function, acting on proximal limb joints' range of motion, and trunk flexibility [264]. They are part of any motor training in the warm-up phase and can be performed even by elderly subjects with mild to moderate disability [264, 265]. Secondly, the selected set of exercises aimed at testing the system on different body segments (arms, trunk, and lower limb) and on the three axial spatial plans assessing. Subjects were asked to repeat each exercise six times, in order both to mimic a real training and obtain an average motor behavior, useful for a reliable assessment.

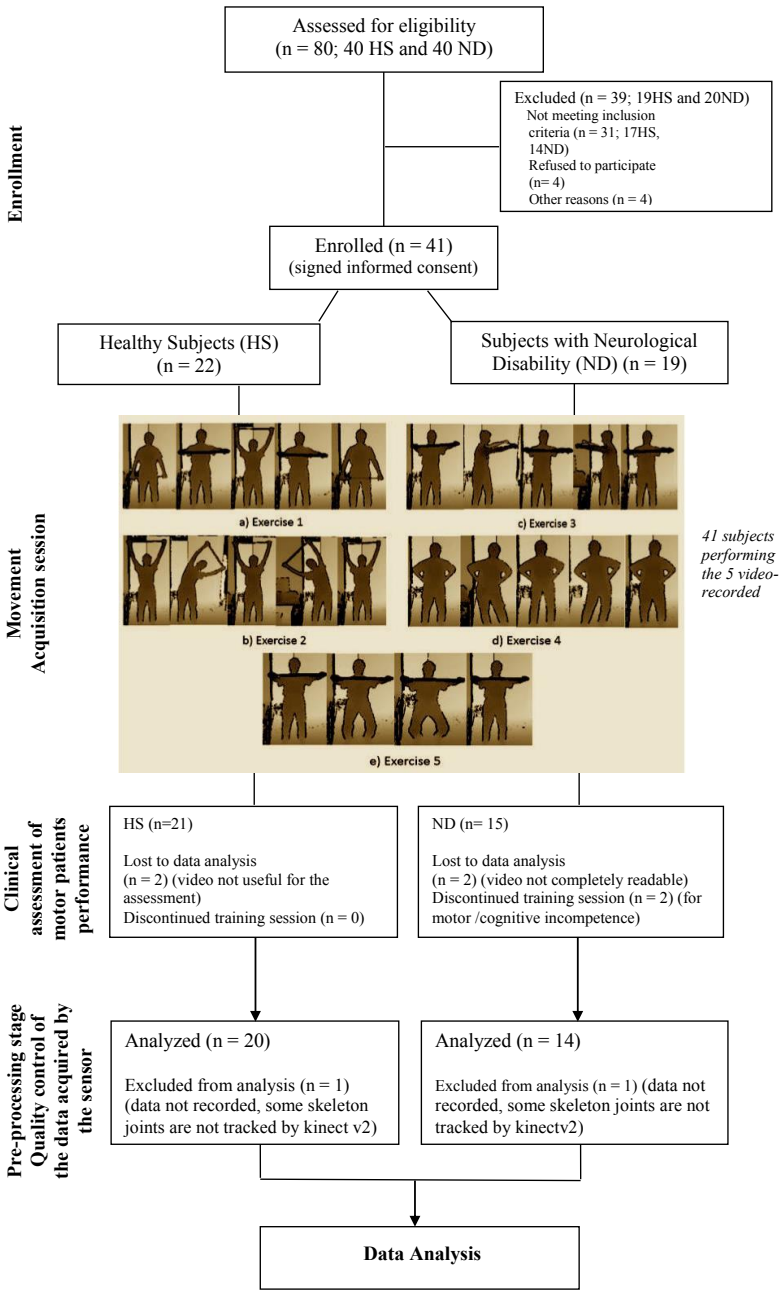
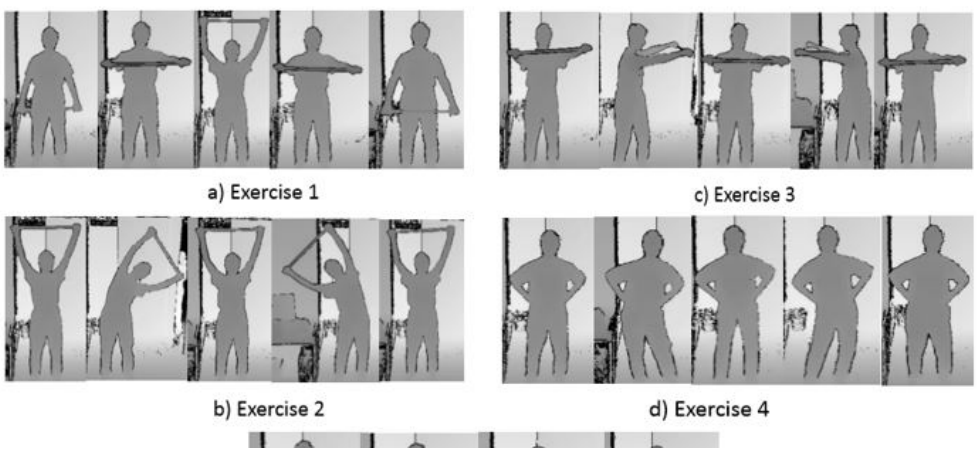


Figure 3.6.: Diagram showing the flow of participants in the study.



3.2. Emotion Inference using Physiological predictors

The benchmark dataset and the collected dataset are described in Section 3.2.1.

3.2.1. Datasets

In order to evaluate the reliability of the proposed methodology, two experiments within the multimedia scenario were performed. The first experiment used high-quality data from the DEAP dataset [138]. Their database includes the possibility to classify 3-D emotion dimensions induced by music video showed to different users. The author aimed at studying how the proposed MIL-based affective state classification methods act when the physiological signal is gathered by accurate sensors in a controlled environment. In this setting, the user is monitored with an accurate set of sensors and the labeling of the perceived emotion has been based on participants' self-reports. The second experiment aimed to investigate the reliability of the method in the real world usage. Data were collected by an unobtrusive smart watch sensor in a less-controlled environment, where the authors in collaboration with psychologists designed the data collection and the labeling procedure. The so-called Consumer dataset aims to provide the kind of data available in the real-world applications, such as the multimedia recommendation system.

DEAP Dataset

In the DEAP Experiment [138], 32 healthy participants watched 40 music videos. The video selection was performed by a semi-automated method, with the main goal of minimizing bias. From the initial candidate of 120 stimuli (60 videos collected from a database and 60 manually collected in order to maximize the clearness of emotional reaction for each of the quadrants), the final 40 test music videos were chosen by using a web-based subjective emotion assessment interface. Each of the 40 resulting videos lasted 1 min. The experiment was performed in two laboratory rooms with controlled illumination. Firstly a 2 minute of baseline stage was recorded, during which a fixation cross was displayed to the participant, who was asked to relax during this period. Then, the 40 videos were presented in 40 trials, each consisting of the following steps:

1. The trial number was displayed to inform the participant of their progress for 2 second.

2. Baseline recordings of 5 seconds (fixation cross).

3. Music video of 1 minutes.

After each video, participants were asked to self-report their emotional experience in four dimensions: *valence*, *arousal*, *dominance* and *liking*. The ratings for *valence*, *arousal*, and *dominance* could range from 1 to 9 and were collected through the Self-Assessment Manikins (SAM)[151]. For the *liking* scale, thumbs down/thumbs up symbols were used. Full-scalp EEG and thirteen peripheral physiological signals including Galvanic Skin Response (GSR), respiration amplitude, Skin Temperature (ST), Blood Volume Pressure by plethysmograph (BVP), electromyograms of Zygomaticus and Trapezius muscles, and electrooculogram (EOG) were recorded at a sampling rate of 512 Hz.

Consumer Dataset

Twenty-nine volunteers with no cognitive diseases and no stress conditions (14 females and 15 males, 20-30 age-range) were recruited at the Department of Information Engineering (Polytechnic University of Marche, Ancona Italy). None of them had a history of neurological disorders. Before the experiment, all participants provided written informed consent. They were asked to watch 6 movie clips (each lasting 4 minutes) and self-reported through the SAM their emotional experience to each video in two dimensions: *valence* and *arousal*. Figure 3.8 shows the flow chart of the experiment setting: tools and instruments displacement, data gathering and storage. The experiment was composed of two different stages performed in different days. Both stages started with a resting stage (baseline stage), where subjects were asked to relax for 10 minutes lying on a couch without sleep. Then, during the first stage, three movie clips were presented to each participant. These videos were chosen for the purpose of eliciting positive *valence* (i.e., happiness, satisfaction). The movie clips proposed in the second stage were instead selected with the goal to elicit negative *valence* (i.e., sadness, fear). At the end of each movie clips, subjects self-reported his/her *valence* and *arousal* level into 9 points SAM scale [151], in a similar way of what was done in [138, 152].

The physiological signals were collected from the smartwatch sensors (Microsoft Band 2 [266]) worn on participants' wrist. A mobile application was implemented to gather the physiological measurement from smartwatch to mobile phone via Bluetooth connection recording data in a .csv file. The data collection was properly synchronized to the movie clips. Three physiological signals were recorded: the Inter-Beat Interval (IBI) of Heart Rate (HR), the Galvanic Skin Response (GSR) sampled at the frequency of 5 Hz and the Skin Temperature (ST) sampled at the frequency of 0.03 Hz.

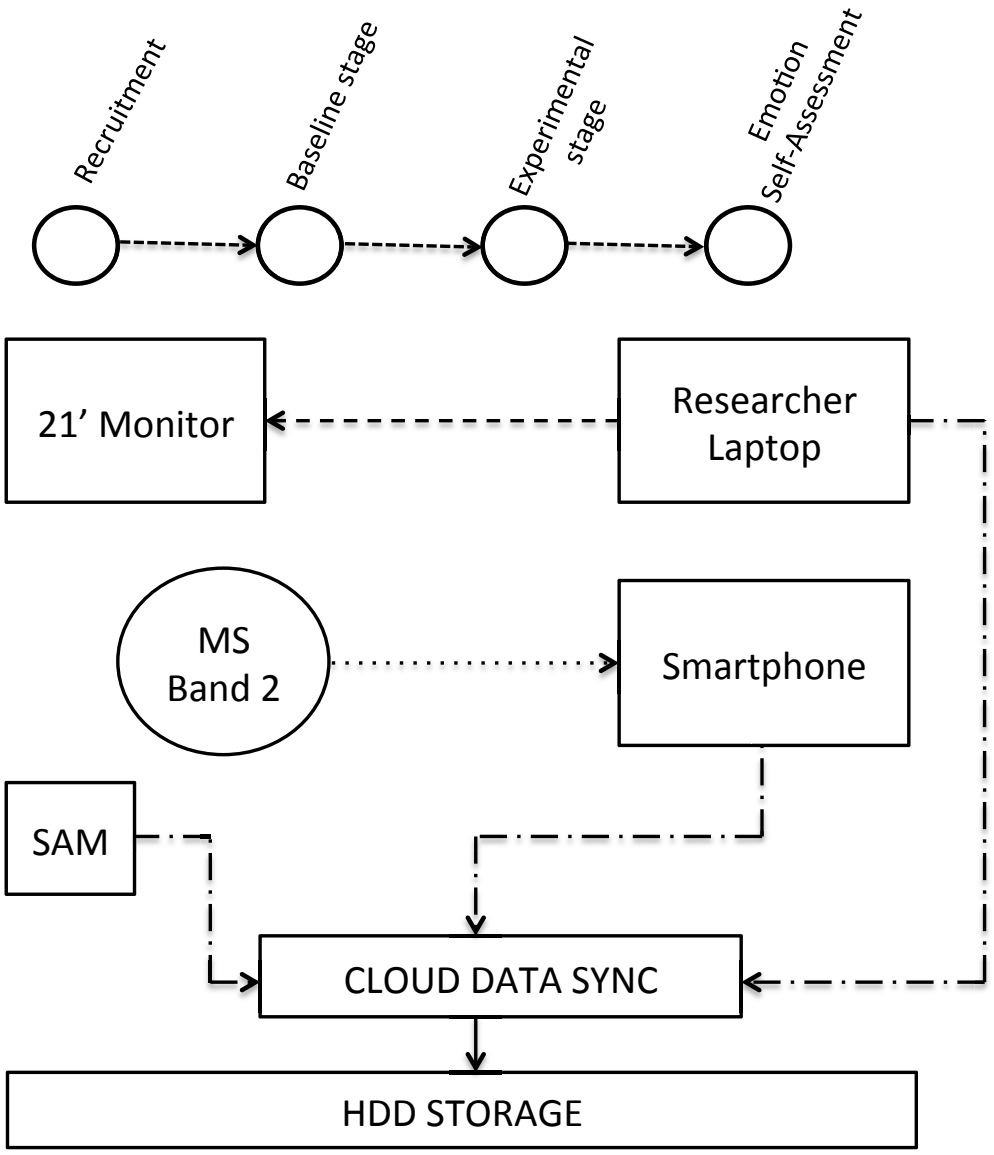


Figure 3.8.: A flow description of the Consumer dataset experiment

Chapter 4.

Method: Computational models

4.1. Quantitative Assessment of Human Movement

For each selected exercises previously described in Section 3.1.3, the clinical assessment is provided (see Section 4.1.1) in order to have a ground truth. Features extraction and segmentation procedure are explained respectively in Section 4.1.2 and Section 4.1.3. Finally, the proposed machine learning method for exercise assessment is described in Section 4.1.4. The presented approach was published in [267].

4.1.1. Clinical Assessment

Two expert clinicians, separately, scrutinized the recorded videos and filled the 10-item Likert questionnaire employed also in [268] and detailed in Appendix A. The questionnaire has been purposely designed to assess the performance of the subject during the exercise. The first three items investigate the accuracy of the exercise targets (i.e., extension of the upper limbs, trunk rotation with upper limbs elevated to 90°, squatting, etc.), while the last seven items evaluate the posture maintained by 7 body segments (head/neck, trunk, arms, pelvis, and legs) during the exercise.

4.1.2. Feature Extraction

The following stage, realized in straight collaboration with clinicians, aims to map the exercise objectives, namely motor functional targets and postural constraints, into kinematic parameters extracted by the 3D joint trajectories. The required features can be divided into Target Features (TFs), Target Velocity Features (TVFs) and Postural Features (PFs). TFs refer to targets to achieve in terms of angles and distances (e.g., maximum knee angle flexion during squatting), TVFs describe the targets in terms of movement speed, while PFs are constraints, angles or distances between anatomical landmarks, that have to be maintained during the exercise (e.g., complete elbow extension during "lifting of the arms" exercise). The accuracy of Kinect v2 in tracking a subset of these features has been reported in Section 2.1.2.

Figure 4.1 shows the extracted features across different body planes (frontal/horizontal) and sides (left/right), as reported also in Table 4.1.

Considering a moderate repetition speed, a pre-processing stage is designed to filter data from artifacts and noise. A second order, zero-phase, low-pass Butterworth filter, with cut-off frequency $f_c = 1$ Hz, is applied for each tracked joint to avoid temporary spikes. For completeness, a brief description of each exercise features has been reported below.

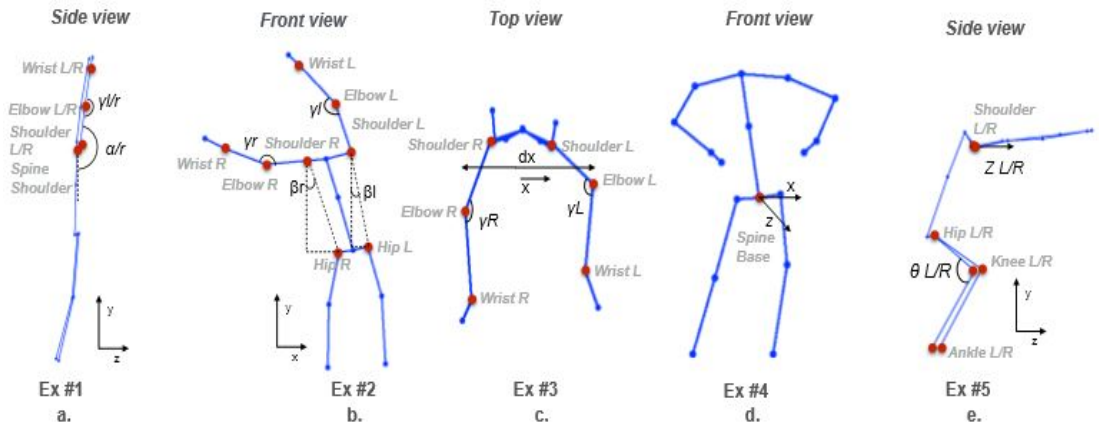


Figure 4.1.: Features extracted for each exercise.

a. **Exercise #1**: right and left armpit angles in sagittal plane ($\alpha_{l/r}$) represent the target movements. Elbow extension angles ($\gamma_{l/r}$) is the constraint to be considered.

b. **Exercise #2**: right and left angles between shoulders and hip ($\beta_{l/r}$) in frontal plane (x,y) are defined as target features, while elbow extension ($\gamma_{l/r}$) is the constraint.

c. **Exercise #3**: target movement is the horizontal distance between elbows (d_x), normalized respect to maximum variation. Elbow extension ($\gamma_{l/r}$) is the constraint.

d. **Exercise #4**: motion is given by spine base trajectories, normalized to zero mean, in transverse plane (x,z), to ensure subject position independence from sensor. This exercise does not contain a PF.

e. **Exercise #5**: right and left knee angles in sagittal plane ($\theta_{l/r}$) are target features. The depth coordinate of the shoulders ($z_{l/r}$) describes constraints, normalized to zero mean.

Table 4.1.: Extracted body features.

Features Extracted	Ex #1	Ex #2	Ex #3	Ex #4	Ex #5
Side	Left/ Right	Left/ Right	Horizontal	Frontal/ Horizontal	Left/ Right
Target Features (TFs)	$\alpha_{l/r}$	$\beta_{l/r}$	d_x	x, z	$\theta_{l/r}$
Target Features Velocity (TVFs)	$\dot{\alpha_{l/r}}$	$\dot{\beta_{l/r}}$	V_{d_x}	V_x, V_z	$\dot{\theta_{l/r}}$
Postural Features (PFs)	$\gamma_{l/r}$	$\gamma_{l/r}$	$\gamma_{l/r}$	\	$z_{l/r}$

4.1.3. Segmentation

Features segmentation aims to locate starting and ending points of each specific movement [91]. The movement is considered as a single repetition performed by the subject. Then, Zero Velocity Crossings (ZVC) [269] is applied to segment each repetition of TF. Among these stationary points, only local minima under specific amplitude and temporal threshold are selected in order to avoid spurious peaks. The amplitude threshold is empirically set as the mean value of the considered feature, while the temporal threshold t_{th} is selected using the recorded samples m and the number of repetitions performed by the subjects (i.e., $R = 6$) as:

$$t_{th} = \frac{m}{2R}. \tag{4.1}$$

The temporal threshold avoids underestimating each segment. Note that the stationary points, obtained by ZVC, are used to segment also TVF and PF. One of the six segmented candidates of TF, TVF and PF are depicted in Figure 4.2 for each Exercise.

4.1.4. Exercise Assessment: HSMM based approach

A HSMM based method is proposed for rehabilitation assessment [262]. Figure 4.3 shows a flow-chart of the proposed algorithm. The training of the Hidden Semi-Markov model is realized using only the features of the Healthy Subject group HS ($n = 7$), who achieved the highest Clinical Score (CS) and the best match between clinician opinions. This group represents the 70% of subjects with a CS higher than 80/100. During the test stage, the assessment score was measured through the log-likelihood. Total and local scores were computed considering the multivariate and univariate observation features vector, respectively.

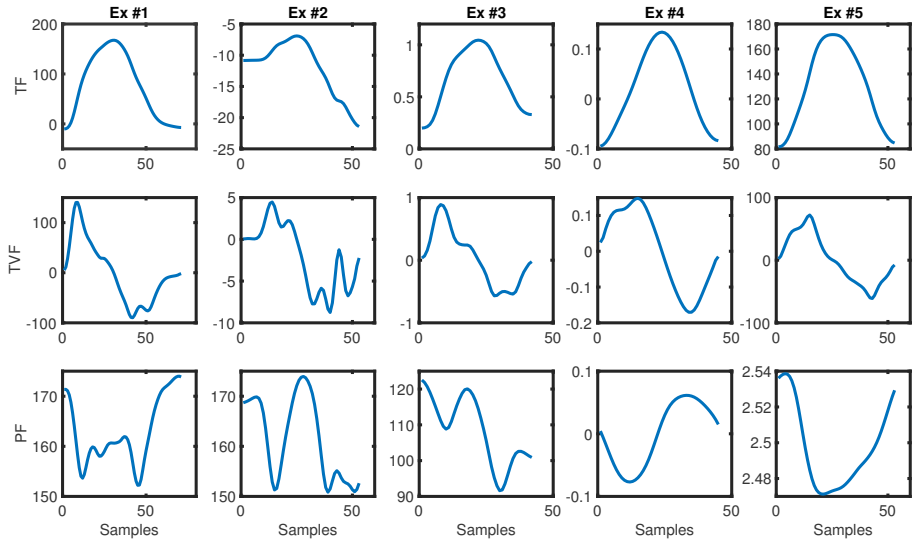


Figure 4.2.: Segmentation examples of TF, TVF and PF for each exercise.

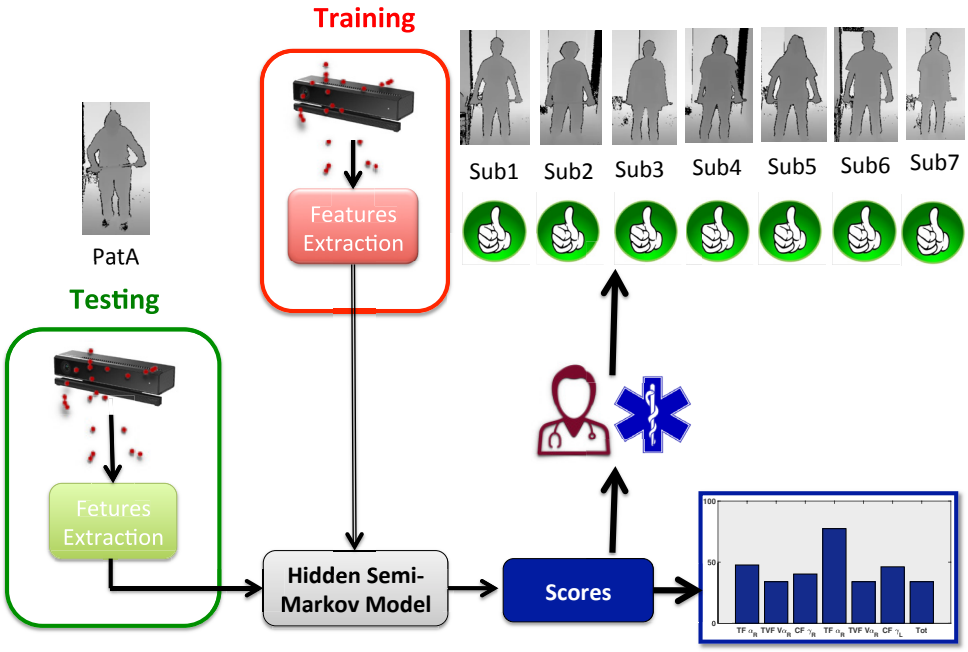


Figure 4.3.: Overview of the proposed algorithm.

HSMM background

HSMM is defined by a discrete-time finite-state homogeneous Markov chain observed through a limited number of transition densities, denoted by the chain states, due to the non-zero probability of self-transition. The state duration of an HMM is implicitly a geometric distribution and one observation per state is assumed [270]. Originating from the popular Hidden Markov Model (HMM) [271], an HSMM is defined expanding the underlying stochastic process to be a semi-Markov chain, where the duration of each state is modeled as a distribution. The following parameters characterize an HSMM: the number of states (ns) in the chain, the duration in the state, limited by the maximum period (σ), the initial state and observation symbol of the probability distribution, and the state transition. The HSMM is used to encapsulate position and velocity information with the parameterization on the involvement of time and space constraint. To calculate the predicted probabilities for the HSMM, the Forward-Backward (FB) algorithm elaborated in [272] is applied. This approach defines the FB variables using the notion of a state together with its remaining sojourn (or residual life) time, the state duration distributions can be taken into account in the Viterbi algorithm where it can be estimated efficiently [273]. Viterbi algorithms are dynamic programming algorithms for the maximum likelihood estimate (MLE) of state sequence of HSMMs [270].

HSMM based assessment

The seven best subjects who reach the highest CS are used to train the HSMM and the segmented TF, TVF and PF represent the observation features vector. During the test stage, the Forward-Backward (FB) and Viterbi algorithms, are applied to compute the filtered, predicted and smoothed probabilities [272, 273]. The HSMM provides a total score, which is computed considering the likelihood function of the multivariate observation sequence as defined in [272], composed by the n-dimensional feature space (i.e., TF, TVF and/or PF):

$$P\left(o_1^T\right) = \left(\prod_{t=1}^{T} r_t\right)^{-1} \tag{4.2}$$

where T is the length of the observation sequence, and $r_t^{-1} = P\left(o_t | o_1^{t-1}\right)$ is the one-step prediction of the observation, with $r_1^{-1} = P\left(o_1\right)$. Hence, the log-likelihood function for each segmented signal is defined by:

$$log\left(P_k\left(o_1^T\right)\right) = -\left(\sum_{t=1}^{T} log(r_{t,k})\right) \tag{4.3}$$

where k is the number of the considered repetitions. Then, the normalized log-likelihood

functions are computed for each repetition (i.e., $k = 1, \ldots, 6$):

$$logL_{norm_k} = \frac{log\left(P_k\left(o_1^T\right)\right)}{T} \tag{4.4}$$

and the total log-likelihood function $logL = \frac{\sum_{k=1}^{6} logL_{norm_k}}{6}$ is the arithmetic mean of the normalized log-likelihood functions $logL_{norm_k}$.

The total score for each $i-th$ subject is computed normalizing the $logL$ within the minimum and maximum CS:

$$score^i = (CS_{max} - CS_{min}) \times \frac{(logL^i - logL_{min})}{logL_{max} - logL_{min}} + CS_{min} \tag{4.5}$$

where CS_{max} and CS_{min} are the maximum and the minimum clinical scores, respectively, while $logL_{max}$ and $logL_{min}$ are the maximum and minimum of the total log-likelihood. In addition, the local scores are extracted from the HSMM, considering the log-likelihood functions of the univariate observation sequences:

$$\begin{aligned} log\left(P_{k_{TF}}\left(o_1^T\right)\right) &= -\left(\sum_{t=1}^{T} log(r_{t,k_{TF}})\right) \\ log\left(P_{k_{TVF}}\left(o_1^T\right)\right) &= -\left(\sum_{t=1}^{T} log(r_{t,k_{TVF}})\right) \\ log\left(P_{k_{PF}}\left(o_1^T\right)\right) &= -\left(\sum_{t=1}^{T} log(r_{t,k_{PF}})\right) \end{aligned} \tag{4.6}$$

Then, the local scores (i.e., $score^i_{TF}$, $score^i_{TVF}$, $score^i_{PF}$) are computed by Eqs. (4.4) and (4.5).

Fig. 4.4 shows the distributions of the Gaussian states for the training subjects of the first exercise.

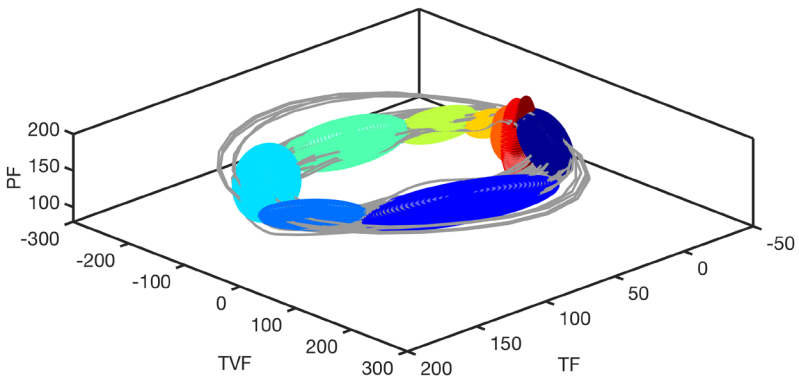


Figure 4.4.: HSMM: distributions of the Gaussian states

4.2. Emotion Inference using Physiological predictors

The features extraction procedure is reported in Section 4.2.1, while the MIL approaches applied for emotion prediction are described in Section 4.2.2.

4.2.1. Features Extraction

The features were extracted as similar as possible for both datasets.

DEAP Dataset

All the acquired physiological measurement were downsampled to 128 Hz and then segmented into 60 seconds trials. The 5 seconds pre-trial baseline was removed and thus not considered for further analyses. The features extraction stage was performed with the Matlab Toolbox for Emotional feAture extraction from Physiological signals (TEAP) [274]. Table 4.2 shows the physiological signals and the resulting 46 features used for the data analysis. Notice that due to the presence of several outliers, the Skin Temperature was not considered in the analysis (see also [274]).

Table 4.2.: Features extracted from physiological signals: DEAP dataset

Signal	Extracted Features
GSR	1) peak amplitude; 2) rising time; 3) number of peaks; 4) average; 5) standard deviation; 6) 1^{st} quartile; 7) 3^{rd} quartile.
Blood volume pressure	8) average; 9) standard deviation of Inter-beat interval (IBI); 10) average of IBI; 11-15) multi-scale-entropy of IBI [139]; 16) spectral power in $0-0.1$ Hz band; 17) spectral power in $0.1-0.2$ Hz band; 18) spectral power in $0.2-0.3$ Hz band; 19) spectral power in $0.3-0.4$ Hz band; 20) energy ratio between the frequency bands $0-0.08$ Hz and $0.15-0.5$ Hz; 21) spectral power of IBI in $0.01-0.08$ Hz band (LF); 22) spectral power of IBI in $0.08-0.15$ Hz band (MF); 23) spectral power of IBI in $0.15-0.5$ Hz band (HF); 24) energy ratio of IBI between MF and LF+HF.
Respiration pattern	25) average; 26) standard deviation; 27) kurtosis; 28) skewness; 29-35) 7 spectral power in the bands from 0 to 2.5 Hz; 36) main frequency.
EMG (2-channels)	37-42) average; 38-43) standard deviation; 39-44) kurtosis; 40-45) skewness; 41-46) spectral power over 20Hz.

Consumer Dataset

The zero-order hold interpolation was applied to resample all physiological signals (i.e., GSR, IBI and ST) at 5Hz. The GSR samples which overcame an empirical

threshold set to 300μS were considered as outliers and deleted using the spline interpolation. All the data were then smoothed using a 5-samples moving average filter. Since the accuracy of the smartwatch data is not comparable with respect to signals gathered from gold standard sensors (see [138]), the extracted features corresponded to a subset of those computed by the major reference works published in the affective computing literature [138, 274, 182, 160] (see Table 4.3). Seven features of IBI $(22-28)$ were computed with respect to values recorded during the baseline stage.

Table 4.3.: Features extracted from physiological signals: Consumer dataset

Signal	Extracted Features
GSR	1) average; 2) standard deviation; 3) average of the derivative; 4) root mean square of the derivative.
IBI	5) standard deviation; 6) standard deviation of the first difference; 7) root mean square of the first differences; 8) number of the absolute values of the first differences samples greater than 50ms (NN50); 9) number of the first differences samples greater than 50ms (dNN50); 10) number of the first differences samples lower than 50ms (aNN50); 11) number of the absolute values of the first differences samples greater than 50ms normalized over the number of samples; 12) average of the absolute values of the first differences; 13) average of the absolute values of the first differences of the normalized signal; 14) average of the absolute values of the second differences; 15) the means of the absolute values of the second differences of the normalized signals; 16) spectral power in $0.04-0.15$ Hz band (LF); 17) spectral power in $0.15-0.4$ Hz band (HF); 18) energy ratio between LF and HF; 19) energy ratio between LF and LF+HF; 20) energy ratio between HF and LF+HF; 21) spectral power in $0.04-0.4$ Hz band (LF+HF); 22) Poincare plot feature $SD1^2$; 23) Poincare plot feature $SD2^2$; 24) Poincare plot feature $SD1^2/SD2^2$; 25) average (IBI^{mean}); 26) $NN50-NN50_{baseline}$; 27) $dNN50-dNN50_{baseline}$; 28) $aNN50-aNN50_{baseline}$; 29) $IBI^{mean}-IBI^{mean}_{baseline}$; 30) standard deviation of the IBI with the mean value calculated from the baseline IBI; 31) $SD1^2/SD1^2_{baseline}$; 32) $SD2^2/SD2^2_{baseline}$;
Skin Temperature	33) average; 34) maximum

4.2.2. Emotion Inference: MIL approaches

In order to formalize the MIL approach and describe the proposed methodology the author uses the following notation:

- $B_1^+,\ldots,B_p^+$ and $B_1^-,\ldots,B_n^-$ are the set of positive and negative training bags, respectively. $\mathscr{B}$ is the set of all such bags.
- x_{ij}^+ and x_{ij}^- are the set of instances in the i-th positive and negative training bag, respectively.
- L is the number of instances for each bag.

The original self-reports of both valence and arousal were binarized by thresholding at level 5 (midpoint). Hence, the author defined as "low", values below 5 and as "high" values above 5, separately for the valence and arousal dimensions. Data processing was implemented to extract salient features for each physiological signal accordingly to the ML model computed. For the standard supervised learning approach each music video/movie clip was considered as a single instance represented by a row-vector of d-features (video-level features). For the MIL models the author evaluated two settings:

- $L = 3$ (i.e., 3 instances x video): each video was segmented in 3 windows overlapped by $ns/4$;

- $L = 5$ (i.e., 5 instances x video): each video was segmented in 5 windows overlapped by $ns/6$.

where ns is the number of recorded samples for each video.

Different MIL approaches were applied. Firstly the EM Diverse Density combined with support vector machine is presented in Section 4.2.2 . Afterwards, the mi-SVM and the MI-SVM approaches are described in Section 4.2.2 and Section 4.2.2 as an evolution of classical single instance learning or normalized set of kernel methods described in Section 4.2.2 .

EMDD-SVM

The key idea behind EMDD is the Diverse Density concept and EM algorithm. Diverse Density (DD) is a measure of the intersection of the positive bags minus the union of the negative bags. Then, by maximizing DD, we can look for both the intersection point (the desired *concept*) and the set of feature weights. Indeed, the DD at a point h in the feature space is a probabilistic measure of both how far the negative instances are from h and how many different positive bags have an instance near c. Formalizing this sentence, the DD of a particular concept h is defined as follows:

$$DD(h) \equiv P(h|\mathscr{B}). \tag{4.7}$$

The application of Bayes Rule leads to find the maximum likelihood estimation:

$$\hat{h} = \arg\max_{h \in H}[P(h|\mathscr{B})] = \arg\max_{h \in H}\left[\frac{P(\mathscr{B}|h)P(h)}{P(\mathscr{B})}\right] \tag{4.8}$$

where H is the hypothesis space. Then assuming independence of the bag instances, uniform prior of the instances and reapplying Bayes rule, leads to:

$$\hat{h} = \arg\max_{h \in H} \left[\prod_{i=1}^{p} P(B_i^+|h) \prod_{i=1}^{n} P(B_i^-|h) \right]$$
$$= \arg\max_{h \in h} \left[\prod_{i=1}^{p} P(h|B_i^+) \prod_{i=1}^{n} P(h|B_i^-) \right]. \quad (4.9)$$

The posterior probability is estimated using the *noisy-or* approximation [275]:

$$P(h|B_i^+) = P(h|x_{i1}^+, \ldots, x_{iL}^+) =$$
$$= 1 - \prod_{j=1}^{L} (1 - P(h|x_{ij}^+)) \quad (4.10)$$
$$P(h|B_i^-) = P(h|x_{i1}^+, \ldots, x_{iL}^-) =$$
$$= \prod_{j=1}^{L} (1 - P(h|x_{ij}^-)) \quad (4.11)$$

where

$$P(h|x_{ij}^+) = \exp(-\|x_{ij}^+ - h\|^2) \quad (4.12)$$
$$P(h|x_{ij}^-) = \exp(-\|x_{ij}^- - h\|^2). \quad (4.13)$$

While learning concept points in the instance space we can also find the best scaling for each k feature that maximizes DD. Then, the Euclidean distance $\|x_{ij} - h\|^2$ becomes the weighted distance $\sum_k s_k^2 \|x_{ij}^{(k)} - h^{(k)}\|^2$. Then, the optimization of DD returns both a location c and a scaling vector s which belong to the hypothesis space H.

The intuition behind EMDD [223] algorithm is to start with some initial guesses of target point h by trying points on the positive bag. Afterwards, in the E-step, the hypothesis h is used to pick one instance from each bag which is most likely to be the one responsible for the label given to the bag. In the second step (M-step), the two-step gradient ascent search [275] of the standard DD algorithm was implemented to find a new h' that maximize $DD(h)$. Since the goal is to classify the video (bag) in the context of affective interaction, the author proposes an adaptation of EMDD algorithm. The final optimal-scaled concept point h is the maximum (in terms of DD) of each scaled concept point computed from each starting instance picked in 10 different positive bags (initial guess). Then, once the scaled concept has been learned, features based on this concept are used to train a linear SVM model. Nearest concept features define the features mapping for each bag as being the minimum distance of any of the instances in that bag to the scaled concepts as follows.

$$\phi(B_i^+) = \min_{1\leq j\leq p}(\sum_k s_k^2 \|x_{ij}^{+\,(k)} - \hat{h}^{(k)}\|^2) \tag{4.14}$$

$$\phi(B_i^-) = \min_{1\leq j\leq n}(\sum_k s_k^2 \|x_{ij}^{-\,(k)} - \hat{h}^{(k)}\|^2). \tag{4.15}$$

The pseudo-code for the proposed EMDD-SVM algorithm is reported in Appendix B (Algorithm 1).

SVM, SIL and NSK

The SVM aims to choose the decision boundary in order to maximize the margin, which is defined to be the smallest distance between the decision boundary and any of the samples. The Single Instance Learning (SIL) is the starting point of SVM approach for MIL: the bag's label is assigned for all instances inside the bag. The optimization problem in SIL is defined as follows:

$$\begin{aligned} \min_{\boldsymbol{w},b,\xi} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\left(\sum_{i=1}^{p}\sum_{j=1}^{L}\xi_{ij}^+ + \sum_{i=1}^{n}\sum_{j=1}^{L}\xi_{ij}^-\right) \\ \text{s.t.} \quad & \\ & (\boldsymbol{w}^T x_{ij}^+ + b) \geq 1 - \xi_{ij}^+ \\ & (\boldsymbol{w}^T x_{ij}^- + b) \leq 1 - \xi_{ij}^- \\ & \xi_{ij}^-, \xi_{ij}^+ \geq 0 \end{aligned} \tag{4.16}$$

where C represents the Box constraint.

The procedure for solving Eq.4.16 is to construct a Lagrange function from the objective function and the corresponding constraints, by introducing a dual set of variables. The key observation is that the Lagrangian solution leads to the dual representation of the maximum margin problem

The optimization of the dual representation subject to a set of inequality constraints takes the form of a quadratic programming problem where the computational complexity in the dual problem depends of the number of instances and bags. If $\hat{\alpha_{ij}}$ is the solution of the dual problem of Eq. 4.16, the prediction of new data points can be computed in terms of the parameters and the kernel function as follows:

$$score_{ij} = \sum_{i=1}^{p+n}\sum_{j=1}^{L}\hat{\alpha_{ij}} y_{ij} K(x\_test_{ij}, x_{ij}) + b \tag{4.17}$$

Note that any instances for which $\hat{\alpha_{ij}} = 0$ will not contribute to the prediction (see Eq 4.17), while the remaining data points constitute the support vectors.

Instead, in the Normalized Set Kernel (NSK) [224] a bag is represented as the sum of all its instances, normalized by its 1 or 2-norm. The resulting representation is used

in training a standard SVM with the following optimization problem:

$$\begin{aligned}
&\min_{\boldsymbol{w},b,\xi} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\left(\sum_{i=1}^{p}\xi_i^+ + \sum_{i=1}^{n}\xi_i^-\right) \\
&\text{s.t.} \\
&\qquad \left(\boldsymbol{w}^T\frac{\sum_j x_{ij}^+}{|B_i^+|} + b\right) \geq 1 - \xi_i^+ \\
&\qquad \left(\boldsymbol{w}^T\frac{\sum_j x_{ij}^-}{|B_i^-|} + b\right) \leq 1 - \xi_i^- \\
&\qquad \xi_i^-, \xi_i^+ \geq 0.
\end{aligned} \tag{4.18}$$

mi-SVM

The maximum pattern margin optimization problem for MIL (mi-SVM, [215]) is defined as follows:

$$\begin{aligned}
&\min_{y_{ij}}\min_{\boldsymbol{w},b,\xi} \quad \frac{1}{2}\|\boldsymbol{w}\|^2 + C\left(\sum_{i=1}^{p}\sum_{j=1}^{L}\xi_{ij}^+ + \sum_{i=1}^{n}\sum_{j=1}^{L}\xi_{ij}^-\right) \\
&\text{s.t.} \\
&\qquad y_{ij}(\boldsymbol{w}^T x_{ij}^+ + b) \geq 1 - \xi_{ij}^+ \\
&\qquad (\boldsymbol{w}^T x_{ij}^- + b) \leq 1 - \xi_{ij}^- \\
&\qquad \xi_{ij}^-, \xi_{ij}^+ \geq 0
\end{aligned} \tag{4.19}$$

where the variable y_{ij} are binary variable associated to the instances in the positive bags, and are bound to satisfy the constrained that $y_{ij} = 1$ for at least one $j \in \{1,\ldots,L\}$.

Notice that the mi-SVM optimization problem (4.19) is a mixed integer programming problem that can only be tackled with heuristic methods. In particular we used the optimization heuristic proposed in [215]. The mi-SVM starts by training a SIL-SVM described above. This is followed by a relabeling of the instances in the positive bags using the SIL decision hyperplane. Hence, if a positive bag contains no instances labeled as positive, the instance that gives the maximum margin of the decision hyperplane is relabeled as positive. This relabeling procedure is repeated, retraining a new SVM model until no labels are changed. The pseudocode is reported in Appendix B (Algorithm 2).

MI-SVM

The maximum bag margin formulation (MI-SVM) is an alternative way of applying the maximum margin approach to the MIL scenario. Since the goal is to have at least one instance in a positive bag to be positive, the aim is to obtain at least one instance in

a positive bag to have a large positive margin. Hence, the maximum over all instance in each positive bag must be bigger than one:

$$\begin{aligned} \min_{\boldsymbol{w},b,\xi} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\left(\sum_{i=1}^{p}\xi_i^+ + \sum_{i=1}^{n}\sum_{j=1}^{L}\xi_{ij}^-\right) \\ \text{s.t.} \quad & \\ & (\boldsymbol{w}^T x_{ij}^- + b) \leq 1 - \xi_{ij}^- \\ & \max_j(\boldsymbol{w}^T x_{ij}^+ + b) \geq 1 - \xi_i^+ \\ & \xi_{ij}^-, \xi_i^+ \geq 0. \end{aligned} \tag{4.20}$$

The second constraint in this optimization problem is not convex. By introducing an extra variable $s(i)$ for each bag, is possible to convert the above formulation into a mixed integer program. Then, in the bag-centered formulation only one pattern per positive bag will determine the margin of the bag. These patterns can be identified as the witness of the entire bag. Hence, the MI-SVM formulation can be resumed by [215]:

$$\begin{aligned} \min_{s}\min_{\boldsymbol{w},b,\xi} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2 + C\left(\sum_{i=1}^{p}\xi_i^+ + \sum_{i=1}^{n}\sum_{j=1}^{L}\xi_{ij}^-\right) \\ \text{s.t.} \quad & \\ & (\boldsymbol{w}^T x_{ij}^- + b) \leq 1 - \xi_{ij}^- \\ & (\boldsymbol{w}^T x_{is(i)}^+ + b) \geq 1 - \xi_i^+ \\ & \xi_{ij}^-, \xi_i^+ \geq 0. \end{aligned} \tag{4.21}$$

Since, as in the mi-SVM, the mixed integer program is difficult to solve for the global optimum, the heuristic algorithm proposed in [215] was employed. The initialization was performed considering the NSK formulation for each positive bag. The witness instance for each positive bag is selected as the instance with the maximum value with respect to the learned decision function. Then, the SVM is retrained with the new dataset, and the procedure is repeated until convergence when the value of s_i stops changing. Note that MI-SVM approach effectively ignores the negative instances in positive bags, and only one instance in the positive bags contributes to the optimization of the hyperplane. On the other hand, in mi-SVM the negative instances in the positive bags, as well as multiple positive instances from one bag can be support vectors. The reader can refer to the pseudocode in Appendix B (Algorithm 3).

Chapter 5.

Results

5.1. Quantitative Assessment of Human Motion

The proposed computational model is evaluated following the data analysis procedure explained in Section 5.1.1. The author provides the results of the proposed HSMM based approach in Section 5.1.2.

5.1.1. Data Analysis

Data analysis was performed in order to:

- provide an inter-rater correlation (r) between the clinical scores assigned through the two compiled questionnaires, computed through Z test. The CS, used for data analysis, was the arithmetic mean of the scores assigned by clinicians, normalized to a 0 - 100 scale;

- measure the Pearson correlation between the proposed method and CS;

- compare the presented method with a DTW algorithm, as proposed in [92, 55]. The scores, computed by both approaches, were correlated with CS;

- test the predictive power of the model, comparing the Root Mean Square Error (RMSE), the Mean Absolute Error (MAE) and the Mean Absolute Percentage Error (MAPE) with respect to CS for each exercise;

- apply the one-way ANOVA. The correlation and ANOVA analysis only included subjects with complete data and statistical significance was set at the .01 level. Moreover, the ROC analysis was performed;

- compare the computation time of the two methods (i.e., HSMM and DTW) for Exercise #1 in order to assess their computation efficiency.

5.1.2. HSMM Results

Thirty-four subjects (17 female; age range: 22 – 76 years) were analysed in the study: twenty HS (11 female; age range: 22 – 50 years) and fourteen (6 female; age range: 30 – 76 years) people suffering from chronic disabilities due to Neurological or musculoskeletal Disabilities. Among ND, six subjects suffered from advanced Parkinson's disease, four from post-stroke hemiparesis and the other from spondylosis. The inter-rater correlation (r), for the CS questionnaire, was as follows: $r = .88$ ($Z = 17.7$; $p < .01$; $[.84 - .91]$) when assessing the total subjects performing all the five exercises, $r = .77$ ($Z = 10.1$; $p < .01$; $[.68 - .84]$) when assessing the HS group, $r = .90$ ($Z = 12.1$; $p < .01$; $[.85 - .94]$) when assessing the ND group.

The hyper-parameters of HSMM were configured maximizing the correlation of HSMM score with respect to CS in the validation set. The considered hyper-parameters are respectively: (1) the Subset of Best Subject $|SBS|$ among the 7 best-performing subjects (bs) identified by CS, (2) the number of hidden states (ns), and the maximum period (σ). The training set is composed by the best training subsets ($SBS \subseteq bs$), while the validation set is represented by the 30% of the $34 - bs$ subjects, the remaining 70% is used for the test set. Then, the HSMM is trained using $C_{n=7}^{|SBS|}$ combinations with a different subset of best subjects, to obtain the best correlation with CS in the validation set. Table 5.1 shows the Pearson correlation (r) between $score_{HSMM}^{i}$ and CS for the test set, considering the best parameters obtained during validation stage.

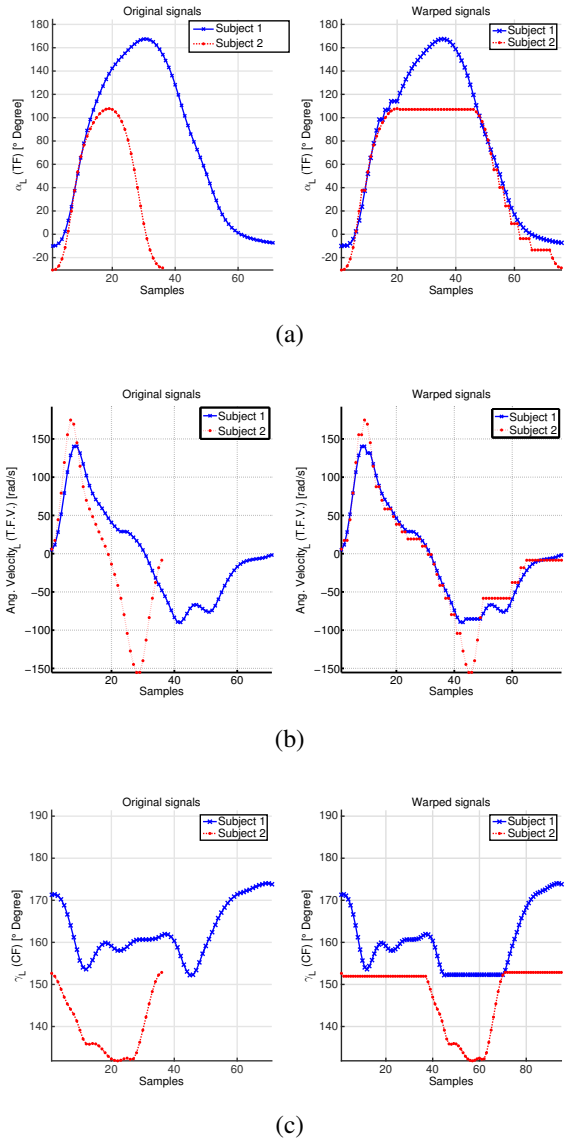
Table 5.1.: Correlation results obtained between $score_{HSMM}$ and CS for the test set with the optimal hyper-parameters.

	Ex. #1	**Ex. #2**	**Ex. #3**	**Ex. #4**	**Ex. #5**
r	.83	.81	.57	.49	.37
p	$< .01$	$< .01$	$< .01$	$< .01$	$< .05$
SBS	2,4	1,2,4	3,4	1,2,6,7	3,4
ns, σ	6,5	6,20	7,20	6,10	9,30

The separate analysis per single exercise showed that the correlation between HSMM scores and CS was high and significant for Exercises #1 and #2. HSMM achieved a moderate and significant correlation for Exercise #3 and #4, while a medium-low and significant correlation was obtained for Exercise #5.

Comparison to DTW

The DTW approach finds an optimal alignment between two time-series, warped in a nonlinear fashion to match each other [276]. This allows evaluating a distance measurement between two features, minimizing the effect of speed variation and time distortion. For instance, in this context, the performance of the subject can be compared to that of a physiotherapist, even though the two sequences have a different duration (see Figure 5.1).



(a)

(b)

(c)

Figure 5.1.: DTW: original signals and warped signals for TF (a) TVF (b) and PF (c)

A common DTW variant, that uses global constraint conditions on the admissible warping paths, is here adopted. Such constraints speed up the DTW computation and prevent pathological alignments by controlling the route of a warping path. Then, a warping path relative to a region R entirely runs within the same region. In this context, the well-known Sakoe-Chiba global constraint region is used [276], and the horizontal and vertical width of the main diagonal (ω) is fixed.

Although the features considered as input to DTW are different with respect to those used in [55] and [92], the approach is the same. Reference time series of a single exercise is composed of features, segmented for each repetition, according to ZVC algorithm. The total DTW distance ($dist$) is the arithmetic mean of the DTW distance computed for TF, TVF and PF averaged over all repetitions. The total score for each

i-th subject is computed normalizing the $dist^i$ within the minimum and maximum CS:

$$score^i_{DTW} = (CS_{min} - CS_{max}) \times \frac{(dist^i - dist_{min})}{dist_{max} - dist_{min}} + CS_{max} \tag{5.1}$$

where $dist^i$ is the arithmetic mean of the relative distances $dist^i_{TF}$, $dist^i_{TVF}$ and $dist^i_{PF}$. Moreover, the local scores (i.e., $score^i_{DTW_{TF}}$, $score^i_{DTW_{TVF}}$ and $score^i_{DTW_{PF}}$) are computed by the Eq. (5.1) associated to TF, TVF and PF for each i-th subject.

In order to compare the performances between the proposed methodology and the DTW approach ([55], [92]), the same validation phase is performed to find the best DTW hyper-parameters (i.e., SBS and (ω)). The correlation and related p-values with respect to CS are shown in Table 5.2, considering all test subjects, ND and HS, respectively.

The correlation coefficient r, obtained for each group, shows that the HSMM based approach outperforms the DTW algorithm for all the proposed exercises except for the Exercise #4. In particular, for the first three exercises, involving upper body movement (i.e., shoulders and arms rotation in the sagittal and transverse plane), HSMM showed a higher correlation with CS respect to DTW, while in Exercise #5 the medium correlation found with HSMM is higher than that obtained by DTW and statistically significant. Moreover, considering only the ND group, the presented approach shows a higher correlation with the clinical assessment respect to DTW for all the exercises, while for HS the correlation is low and not statistically significant for both methods.

The correlation between HSMM and DTW results high and significant for Exercises #1, #2 and #4, moderate and significant for Exercise #3, and low and not significant for Exercise #5. Furthermore, the outcomes of both approaches for ND are strongly correlated for Exercises from #1 to #4 and medium correlated for Exercise #5, while for HS they show a moderate correlation for all the exercises except for the Exercise #2 and #3.

The predictive power of the model is shown in Table 5.3. For Exercises #1 and #2, the comparison of errors confirms that HSMM is more accurate than DTW, while for Exercise #4, DTW provides better results particularly supported by the MAPE value which is considerably lower. In Exercises #3 and #5 the RMSE values of HSMM are lower than DTW but not for the MAPE values which result higher.

In addition to the total score, authors provide local scores for the involved features that allow the clinician to localize the error in the exercise movement execution. Figure 5.2 shows the local scores, computed by Eq.(4.6), for the physiotherapist (Subject 2) and for a patient (Subject 22) for Exercise #1, by HSMM and DTW. Scores for the healthy subject obtained with HSMM vary from a value of 70 to almost 100, while for the patient they range from 30 to 70. The $score_{TF}$ shows as the patient does not follow the target feature related to the left underarm angle in the sagittal plane (i.e., α_l, as described in Section 4.1.2, Exercise #1) while the right underarm angle (i.e., α_r) shows an higher score (close to 60). The local $score_{TF}$ assigned to the patient is lower

Table 5.2.: Correlation results: HSMM vs CS, DTW vs CS and HSMM vs DTW.

			Ex. #1	Ex. #2	Ex. #3	Ex. #4	Ex. #5
All Test Subjects	**HSMM**	r	**.83**	**.81**	**.57**	.49	**.37**
		p	$< .1$	$< .01$	$< .01$	$< .01$	$< .05$
		Z	6.6	4.4	3.3	2.9	2.2
		$C.I.(95\%)$	.69-.91	.53-.84	.24-.74	.17-.71	.04-.63
		SBS	2,4	1,2,4	3,4	1,2,6,7	3,4
		ns, σ	6,5	6,20	7,20	6,10	9,30
	DTW	r	.70	.68	.52	**.62**	.27
		p	$< .01$	$< .01$	$< .01$	$< .01$	n.s.
		Z	4.6	4.4	3.4	4.0	1.5
		$C.I.(95\%)$	.47-.84	.43-.83	.25-.74	.35-.80	-.07-.56
		SBS	3,4	3	6	3	3,5,6
		ω	5	15	30	10	15
	HSMM vs DTW	r	.78	.91	.66	.71	.30
		p	$< .01$	$< .01$	$< .01$	$< .01$	n.s.
		Z	5.8	8.1	4.3	4.6	1.7
		$C.I.(95\%)$	.60-.89	.82-.96	.41-.92	.48-.85	-.04-.58
ND	**HSMM**	r	**.91**	**.60**	**.56**	**.68**	**.50**
		p	$< .01$	$< .05$	$< .05$	$< .01$	n.s.
		Z	5.10	2.31	2.09	2.59	1.82
		$C.I.(95\%)$	.74-.97	.11-.86	.04-.84	.20-.89	-.42-.82
	DTW	r	.85	.59	.51	.67	.39
		p	$< .01$	$< .05$	n.s.	$< .01$	n.s.
		Z	4.12	2.27	1.85	2.54	1.36
		$C.I.(95\%)$	.57-.95	.09-.86	-.03-.82	.18-.89	-.18-.76
	HSMM vs DTW	r	.83	.96	.77	.88	.54
		p	$< .01$	$< .01$	$< .01$	$< .01$	$< .05$
		Z	3.96	6.3	3.39	4.41	2.46
		$C.I.(95\%)$	.54-.95	.86-.99	.41-.92	.65-.97	.12-.79
HS	**HSMM**	r	.31	-.18	$-.10$	$-.28$	.16
		p	n.s.	n.s.	n.s.	n.s.	n.s.
		Z	1.27	-.69	-.03	-1.19	.64
		$C.I.(95\%)$	-.17-.67	-.61-.33	-.45-.44	-.64-.19	-.31-.56
	DTW	r	.14	.25	.01	.08	-.09
		p	n.s.	n.s.	n.s.	n.s.	n.s.
		Z	.58	.95	.04	.32	-.38
		$C.I.(95\%)$	-.33-.56	-.27-.65	-.43-.45	-.38-.5	-.51-.37
	HSMM vs DTW	r	.59	.25	-.07	.67	.54
		p	$< .01$	n.s.	n.s.	n.s.	$< .05$
		Z	2.74	.95	-.3	1.86	2.46
		$C.I.(95\%)$	.19-.83	-.27-.65	-.5-.38	-.03-.73	.12-.79

than the score assigned to the physiotherapist. Accordingly, the patient does not fully achieve a correct speed of movement (i.e $score_{TVF}$). Moreover, $score_{PF}$, related to postural features, discloses as the patient does not respect almost all constraints (i.e., elbow stretched during the exercises: $\gamma_{l/r}$). The local scores, obtained by DTW, show

Table 5.3.: RMSE, MAE and MAPE values related to HSMM and DTW scores with respect to CS for all test subjects.

	Ex. #1	Ex. #2	Ex. #3	Ex. #4	Ex. #5
HSMM					
RMSE	8.00	13.45	15.32	21.21	22.64
MAE	6.01	10.17	12.35	17.35	18.15
MAPE (%)	3.8	1.87	17.6	50.7	46.8
DTW					
RMSE	11.28	19.02	16.21	14.6	23.57
MAE	7.68	12.03	10.87	11.22	18.54
MAPE (%)	5.7	23.1	10.5	2.77	11.05

a similar trend compared to those obtained by HSMM except that for $score_{DTW_{TF}}$ related to the right underarm angle that was lower, and for the $score_{DTW_{PF}}$ related to right elbow angle that was higher. In addition, DTW underestimates the score related to TVF. On the basis of these discrepancies, clinicians decided to evaluate again the video and the new clinic evaluation agrees with results obtained by HSMM, revealing that the patient did not reach the primary target of the exercise (especially for the left underarm angle), and he/she did not respect the postural constraints.

Discrimination between healthy and disabled people

In order to test the ability of the proposed approach of discriminating between healthy and disabled people, the results, obtained by the three different assessment methods applying the descriptive statistic analysis, are shown in Table 5.4 and Figure 5.3.

Table 5.4.: Descriptive statistics of the scores obtained through the three different assessment methods.

		CS	HSMM	DTW
ND	**Mean ±SD**	68.5 (±23.3)	68.2 (±20.7)	77.2 (±19.9)
	Med(IQR)	72.5 (15)	71.2 (35.8)	84.5 (21.4)
	min-max	0 – 100	26 – 97	26 – 99.5
HS	**Mean ±SD**	87.5 (±15.3)	81.0 (±14)	89.6 (±9.5)
	Med(IQR)	91 (15)	80.8 (25.9)	92.2 (9.8)
	min-max	0 – 100	41.6 – 100	48.1 – 100

DTW algorithm tends to overestimate the clinical assessment for both ND and HS, while the HSMM approach is more efficient for ND respect to HS, for which it underestimates the scores.

The quantitative inter-group comparison, by one-way ANOVA, highlighted that the three methods, CS ($F(2,33) = 40.5,\ p < .01$), HSMM ($F(2,33) = 22.7,\ p < .01$) and DTW ($F(2,33) = 28.2,\ p < .01$) were able to discriminate between patients and

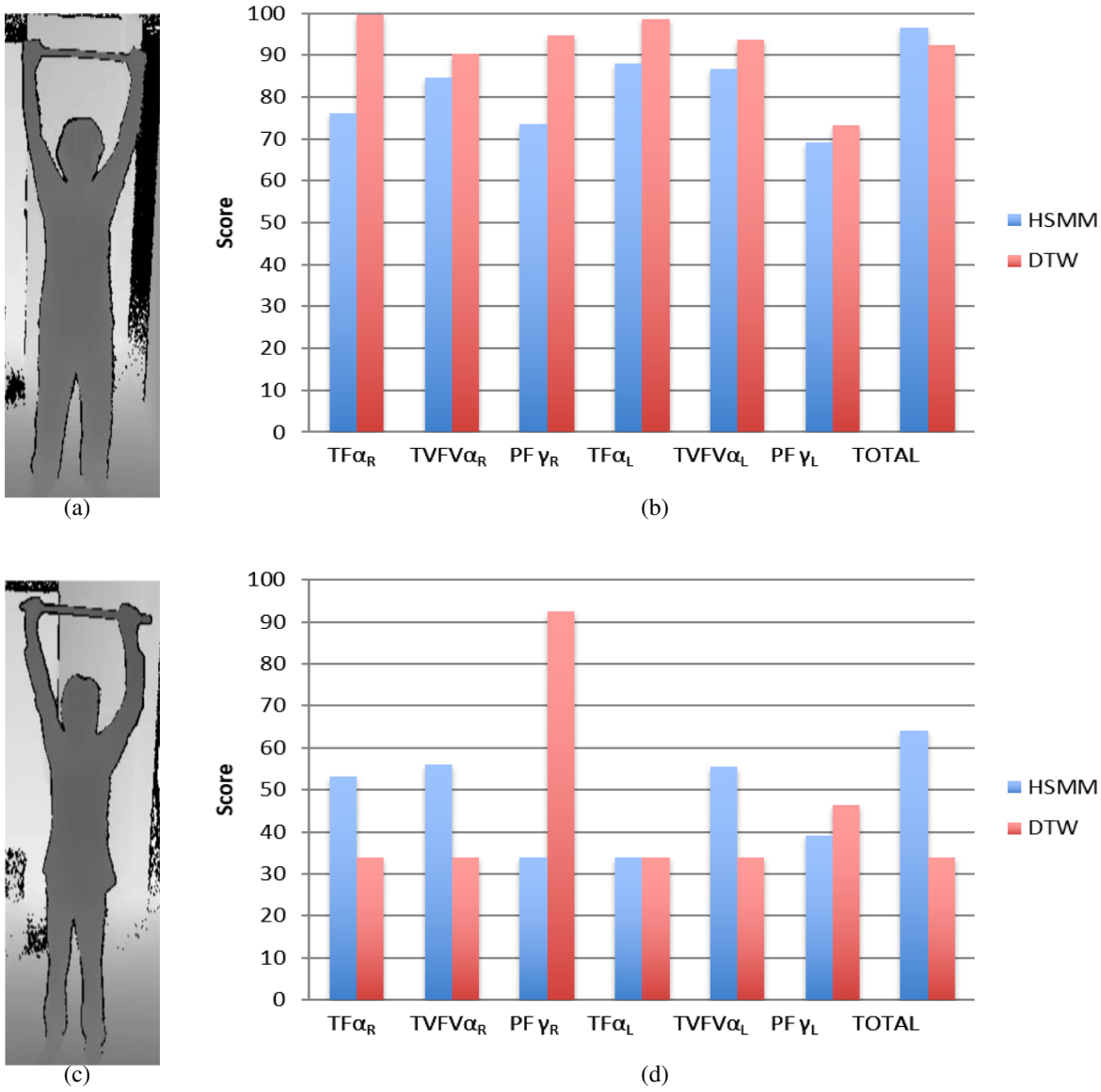


Figure 5.2.: Local and global scores computed by HSMM (blue bar) and DTW (red bar) during Exercise #1: physiotherapist (a-b) vs patient (c-d). A screenshot of the depth image recorded by Kinect v2 is also shown.

healthy subjects.

Figure 5.4 shows the Receiving Operating Characteristic (ROC) curve of HSMM, DTW and CS for discriminating the two groups (i.e., HS and ND). The Area Under Curve (AUC) of CS is greater (AUC=0.755) than DTW (AUC=0.714) and HSMM (AUC=0.681), respectively.

Computation time

Table 5.5 shows the computation time, expressed in seconds (s), for training and validation stage of Exercise #1 averaging over all combinations of $|SBS|$. Both meth-

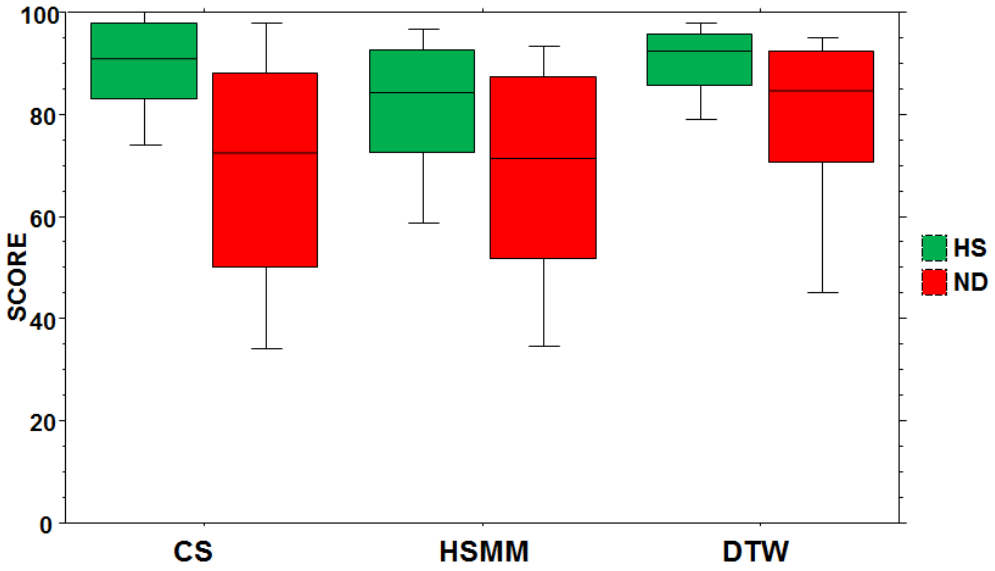


Figure 5.3.: Box plot about intergroup comparison.

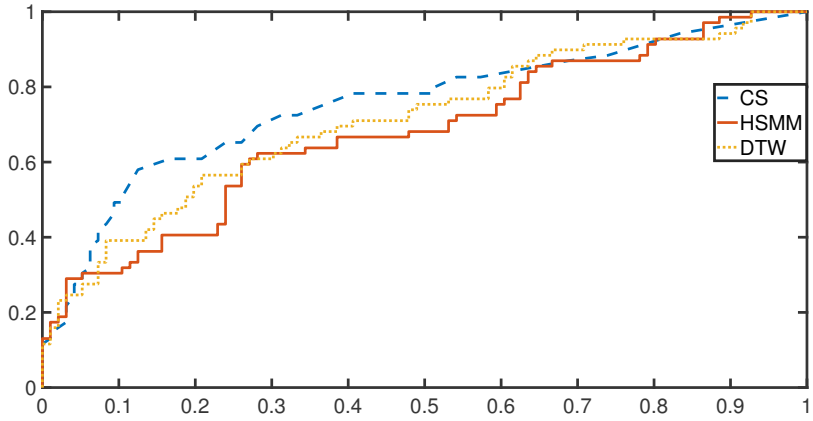


Figure 5.4.: AUC for the three assessment methodologies.

ods (i.e., HSMM and DTW) are reasonably fast and would be practically feasible for motion assessment. HSMM seems to outperform DTW in terms of computation efficiency for each training set size. Accordingly, the gap between HSMM and DTW increases with the number of best subjects considered for the training as in Figure 5.5.

Table 5.5.: Computational time (s) for training and validation.

	$\|SBS\|$						
	1	2	3	4	5	6	7
HSMM	110.96	145.42	199.82	215.59	267.16	329.85	388.46
DTW	118.36	172	235.69	273.28	335.30	441.27	511.41

All the experiments were performed using Matlab 2016*b* on the computer with 16

Gb RAM and *i*5 CPU.

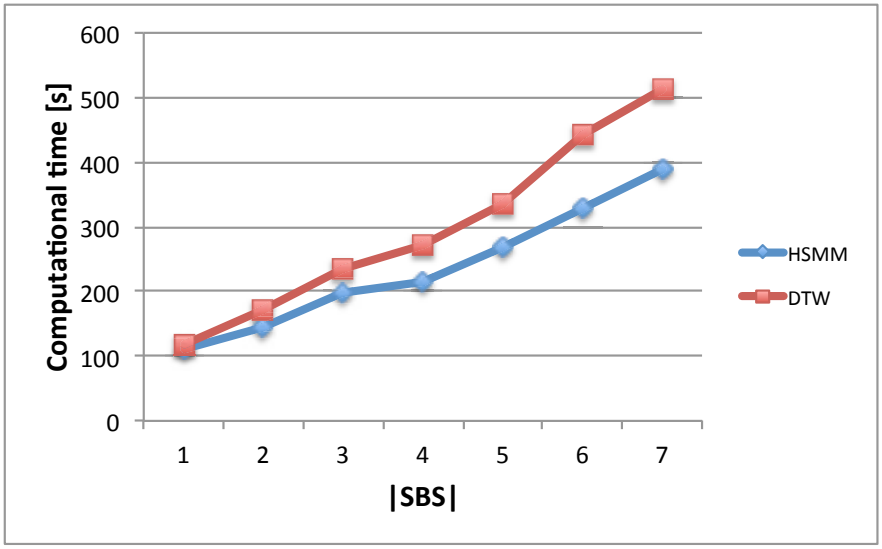


Figure 5.5.: Computation time for training and validation related to the Exercise #1.

5.2. Emotion Inference from Physiological predictors

The proposed MIL framework is evaluated following the data analysis procedure explained in Section 5.2.1. The author provides the results of the proposed MIL based approaches in Section 5.2.2.

5.2.1. Data analysis

The extracted features were modeled by the MIL algorithms described in Section 4.2.2 in order to predict low/high levels of both valence and arousal. The assessment of the MIL model was performed according to the following measures:

- *accuracy*: the percentage of correct predictions;
- *confusion matrix*: square matrix that shows the type of errors in a supervised binary paradigm;
- *macro-F1 score (macro-F1)*: the harmonic mean of precision and recall averaged over all output categories;
- *Receiver Operating Characteristic*: is designed by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. It illustrates the performance of a binary classifier as its discrimination threshold is varied.

The MIL approach was tested on the DEAP dataset building an *user-specific* model with a Leave-One-Video-Out (LOVO) procedure. This is the same testing performed in [138]. Accordingly, in the Consumer dataset the small amount of video samples for each subject did not allow to perform the same analysis. Hence, a 10-fold Cross Validation (10-CV) over video procedure was computed.

The Naive Bayes (NB), the Support Vector Machine (SVM) and the Random Forest (RF) classifiers are standard supervised ML algorithm used as comparison. The NB was employed in [138], while the RF achieved the best performance among other tested supervised classifiers such as Decision Tree and K-Nearest Neighbors. Additionally, from the machine learning point of view, the applied MIL methods are a theoretical extension of SVM.

5.2.2. MIL Results

The linear Kernel is employed in the SVM, EMDD-SVM, mi-SVM, and MI-SVM. The optimization of hyperparameter (i.e., Box Constraint) was performed implementing a grid-search and optimizing the *macro-*F_1 score (F_1) in the validation set. The Box Constraint was picked inside the subset $\{0.1, 0.5, 1, 5, 25, 100\}$.

Experiment 1: DEAP Dataset

Table 5.6 shows the average accuracies and *macro-F1* scores of *user-specific* setup (LOVO) over participants for the standard machine learning algorithm and the MIL methods for each task (i.e., *arousal*, *valence* and *dominance*) and settings (i.e., $L = 3$ and $L = 5$). Concerning the estimation of the two *arousal* level the mi-SVM method with $L = 3$ shows the highest *macro-F1* (*macro-F1*=0.546, ACC=0.611), while the EMDD-SVM approach has the lowest *macro-F1* (*macro-F1*=0.463, ACC=0.559). The MI-SVM method with $L = 5$ has the best outcome (*macro-F1*=0.612, ACC=0.636) for recognizing the *valence* level, while the EMDD-SVM approach has the worst performance (*macro-F1*=0.526, ACC=0.566). The prediction of *dominance* receives relatively low *macro-F1* for both baseline and MIL approaches.

To test for significance, an independent one-sample t-test was performed comparing the *macro-F1* distribution over participants with respect to chance level (.5) (see Table 5.6). For the *valence* task, the *macro-F1* is significantly higher ($p < .05$) than chance level (.5) for both standard and MIL methods. The *macro-F1* obtained from EMDD-SVM with $L = 5$ did not reach significance. On the other hand, for the estimation of *arousal* level only the *macro-F1* of mi-SVM approach with $L = 3$ is significantly higher ($p < .05$) than chance level (.5). Neither of the methods showed *macro-F1* above chance for the prediction of *dominance* level.

Since all the MIL methods are based on SVM approach, we chose to compare the *macro-F1* score, the confusion matrix and the Receiver Operating Characteristic (ROC) of the best MIL approach with those obtained from standard SVM for both the *arousal* and *valence* tasks. Since no significant results were found in the classification of *dominance*, no further analysis was performed for this task.

Figure 5.6 shows the *macro-F1* scores of mi-SVM with 3 windows per video and the standard SVM approach for each participant for the *arousal* task. When compared to standard SVM, the mi-SVM with $L = 3$ shows higher *macro-F1* scores in 18 out of 32 participants (i.e., participant 2, 5-12, 16, 18-19, 22-24, 28-30).

Table 5.7 shows the confusion matrices of the mi-SVM and the standard SVM approach over all participants for the *arousal* task. Notice that the *true high* rate in mi-SVM (0.71) is higher than SVM (0.64), while the *false high* rate in SVM (0.48) is lower than mi-SVM (0.53).

The ROC for mi-SVM and SVM is depicted in Figure 5.7. The Area Under Curve (AUC) of mi-SVM (AUC=0.638) is comparable with this obtained by SVM standard method (AUC=0.636).

For what concerns the estimation of the *valence* level, the performance of MI-SVM with $L = 5$ is higher than SVM method for 22/32 participants (i.e., participant 2, 3, 4, 8-11,13-17, 19-23, 25, 27, 29-31) (see Figure 5.8).

Table 5.8 shows the confusion matrices of the MI-SVM and the standard SVM approach over all participants for the *valence* task. The MI-SVM discloses an higher *true negative* and *true positive* rate (0.58 and 0.68) than SVM (0.51 and 0.64).

Table 5.6.: Average accuracies (ACC) and *macro-F1* (F1) of *user-specific* setup (LOVO) over participants for the MIL algorithms. For comparison, standard results are given for classification based on NB and SVM. Stars indicate whether the *macro-F1* distribution over subjects is significantly higher than chance level (i.e., *macro-F1*=0.5) according to an independent one-sample t-test ($** = p < .01, * = p < .05$)

	Arousal		**Valence**		**Dominance**	
Algorithm	ACC	F1	ACC	F1	ACC	F1
Standard ML						
NB	0.572	0.514	0.595	0.577**	0.566	0.506
SVM	0.591	0.539	0.581	0.557**	0.599	0.530
RF	0.586	0.498	0.599	0.566**		
MIL						
3 windows						
mi-SVM	0.611	0.546*	0.622	0.595**	0.603	0.528
MI-SVM	0.594	0.533	0.577	0.556*	0.573	0.508
EMDD-SVM	0.559	0.463	0.587	0.544*	0.580	0.485
5 windows						
mi-SVM	0.583	0.512	0.621	0.593**	0.593	0.522
MI-SVM	0.585	0.530	0.636	0.612**	0.578	0.518
EMDD-SVM	0.589	0.501	0.566	0.526	0.552	0.455

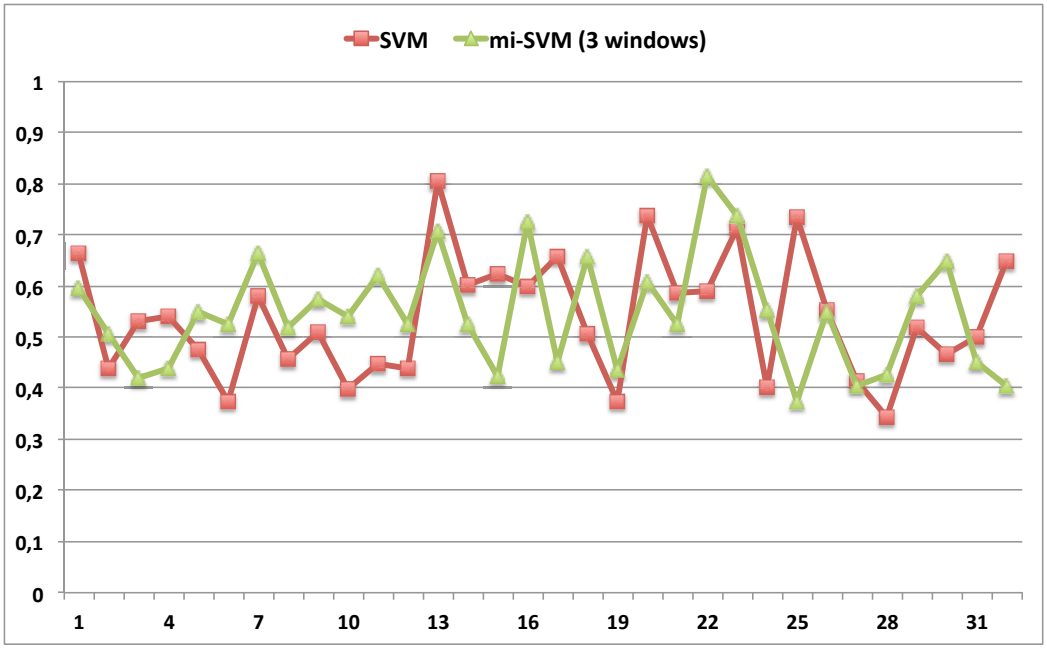


Figure 5.6.: The *macro-F1* score of the mi-SVM with $L = 3$ and the standard SVM approach for each participant for the *arousal* task

Table 5.7.: Confusion matrices (rows are the true classes) of the mi-SVM with $L = 3$ and the standard SVM approach over all participants for the *arousal* task

mi-SVM

	low	*high*
low	0.47	0.53
high	0.29	0.71

SVM

	low	*high*
low	0.52	0.48
high	0.36	0.64

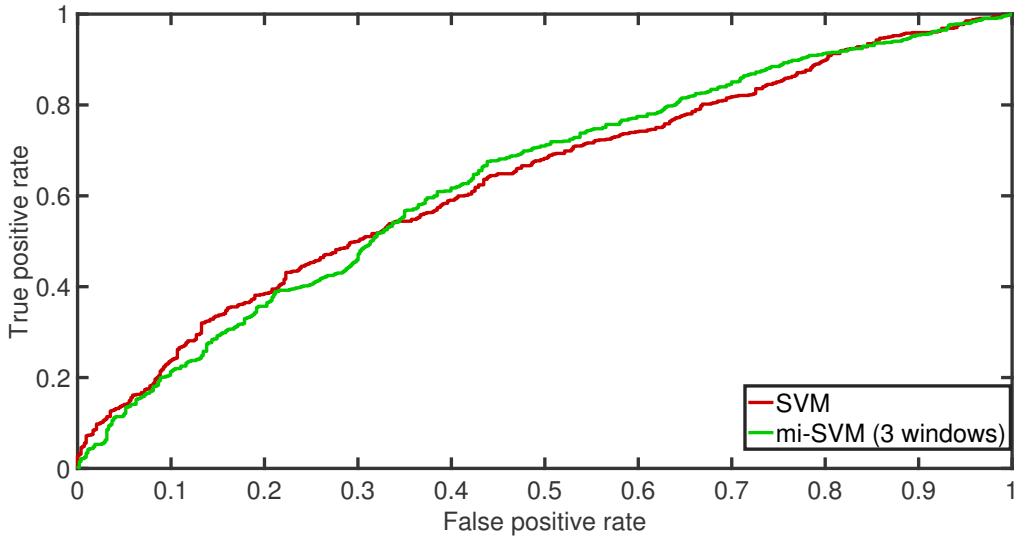


Figure 5.7.: ROC curve of the mi-SVM with $L = 3$ and the standard SVM approach over all participants for the *arousal* task

Table 5.8.: Confusion matrices (rows are the true classes) of the MI-SVM with $L = 5$ and the standard SVM approach over all participants for the *valence* task

MI-SVM

	neg	*pos*
neg	0.58	0.42
pos	0.32	0.68

SVM

	neg	*pos*
neg	0.51	0.49
pos	0.36	0.64

Figure 5.9 shows the ROC for MI-SVM and the standard SVM method. The AUC of MI-SVM (AUC=0.660) is higher than the SVM method (AUC=0.590).

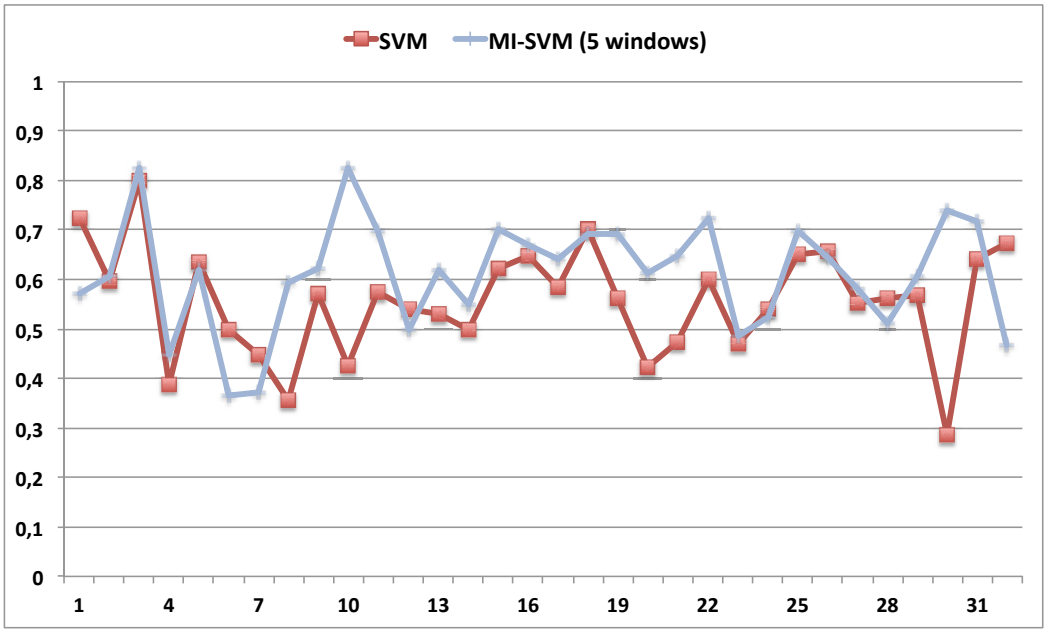


Figure 5.8.: The *macro-F1* score of the MI-SVM with $L = 5$ and the standard SVM approach for each participant for the *valence* task

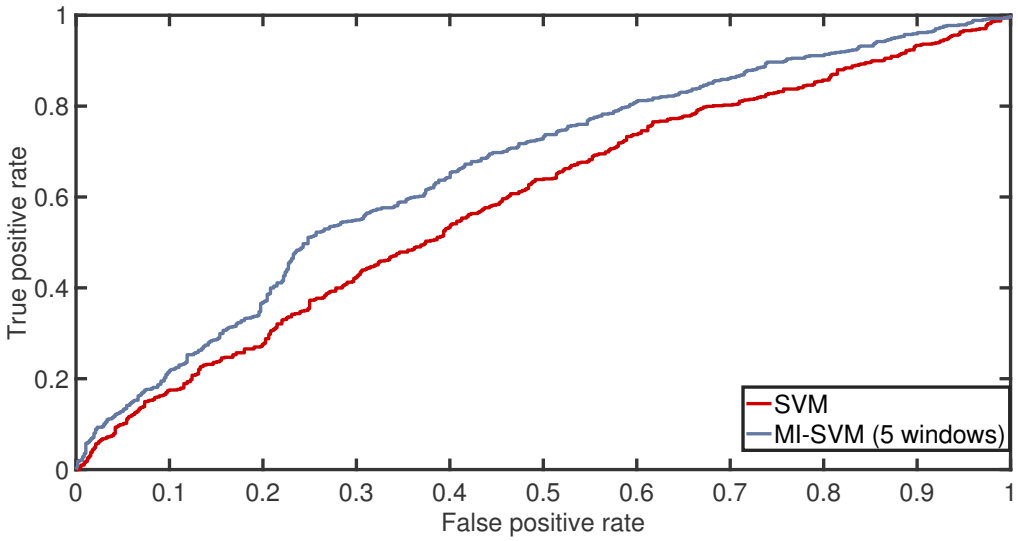


Figure 5.9.: ROC curve of the MI-SVM with $L = 5$ and the standard SVM approach over all participants for the *valence* task

Experiment 2: Consumer Dataset

Table 5.9 shows the average accuracies and *macro-F1* scores of *user-specific* setup (10-CV) for the MIL algorithm and the standard methods for the two dimensions emotional state: *arousal* and *valence*. For the classification of *arousal* level the EMDD-SVM approach with $L = 5$ reveals the highest *macro-F1* (*macro-F1*=0.656, ACC=0.733), while the NB classifier and the EMDD-SVM model with $L = 3$ has the

lowest *macro-F1* (respectively, *macro-F1*=0.429, ACC=0.456 and *macro-F1*=0.428, ACC=0.629). The mi-SVM method with $L = 3$ shows the best performance (*macro-F1*=0.788, ACC=0.789) for the *valence* classification, while the EMDD-SVM approach with $L = 5$ has the worst outcome (*macro-F1*=0.530, ACC=0.537).

Table 5.9.: Average accuracies (ACC) and *macro-F1* (F1) of *user-specific* setup (10-CV) over 10-fold for the MIL algorithms. For comparison, standard results are given for classification based on NB and SVM. Stars indicate whether the *macro-F1* distribution over the 10-fold is significantly higher than chance level (i.e., *macro-F1*=0.5) according to an independent one-sample t-test ($** = p < .01, * = p < .05$)

	Arousal		**Valence**	
Algorithm	ACC	F1	ACC	F1
Standard ML				
NB	0.456	0.429	0.588	0.577
SVM	0.629	0.497	0.651	0.646$*$
RF	0.655	0.598	0.771	0.775$*$
MIL				
3 windows				
mi-SVM	0.644	0.569	0.789	0.788$**$
MI-SVM	0.690	0.528	0.745	0.742$**$
EMDD-SVM	0.629	0.428	0.717	0.713$**$
5 windows				
mi-SVM	0.647	0.535	0.704	0.691$*$
MI-SVM	0.677	0.586	0.688	0.681$**$
EMDD-SVM	0.733	0.656$*$	0.537	0.530

For what concerns the classification of *valence* level, all the MIL models except the EMDD-SVM with $L = 5$ show a *macro-F1* significantly higher ($p < .05$) than chance level (.5), while for the standard methods (i.e., NB and SVM), only the *macro-F1* of the standard SVM overcomes significantly ($p < .05$) the chance level (.5). On the other hand, for the estimation of *arousal*, only the EMDD-SVM with $L = 5$ reveals a *macro-F1* significantly higher ($p < .05$) than chance level (.5).

Figure 5.10 shows the *macro-F1* scores of each fold for both the EMDD-SVM with $L = 5$ and the standard SVM method. The classification of *arousal* level is identical for both methods for in fold 1 and fold 2, while the EMDD-SVM is superior than SVM for 7 out of 10 folds (i.e., 3-6 and 8-10).

Table 5.10 shows the confusion matrices of the EMDD-SVM with $L = 5$ and the standard SVM approach over all folds for the *arousal* task. Both the *true low* and *true high* rate in EMDD-SVM (0.43 and 0.87) is higher than SVM (0.2 and 0.83).

The ROC for EMDD-SVM with $L = 5$ and SVM is depicted in Figure 5.11. The

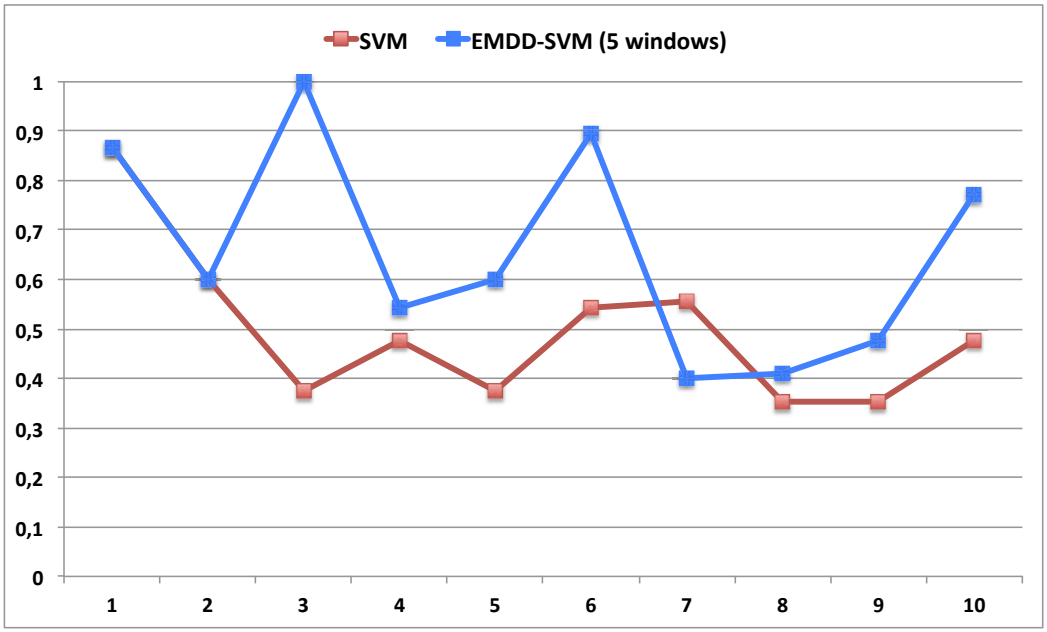


Figure 5.10.: The *macro-F1* score of the EMDD-SVM with $L = 5$ and the standard SVM approach for each participant for the *arousal* task

Table 5.10.: Confusion matrices (rows are the true classes) of the EMDD-SVM with $L = 5$ and the standard SVM approach over all participants for the *arousal* task

EMDD-SVM

	low	*high*
low	0.43	0.57
high	0.13	0.87

SVM

	low	*high*
low	0.2	0.8
high	0.17	0.83

Area Under Curve (AUC) of EMDD-SVM (AUC=0.610) is higher than the SVM (AUC=0.540).

For the *valence* task, compared to standard SVM the mi-SVM with $L = 3$ yields an higher *macro-F1* in 8 out of 10 folds (i.e., fold 2-4 and 6-10) (see Figure 5.12).

Table 5.11 shows the confusion matrices of the mi-SVM and the standard SVM approach over all participants for the *valence* task. The mi-SVM shows an higher *true negative* and *true positive* rate (0.84 and 0.75) than SVM (0.59 and 0.71).

Figure 5.13 shows the ROC curves for mi-SVM and the standard SVM method. The AUC of mi-SVM (AUC=0.829) is higher than the SVM method (AUC=0.729).

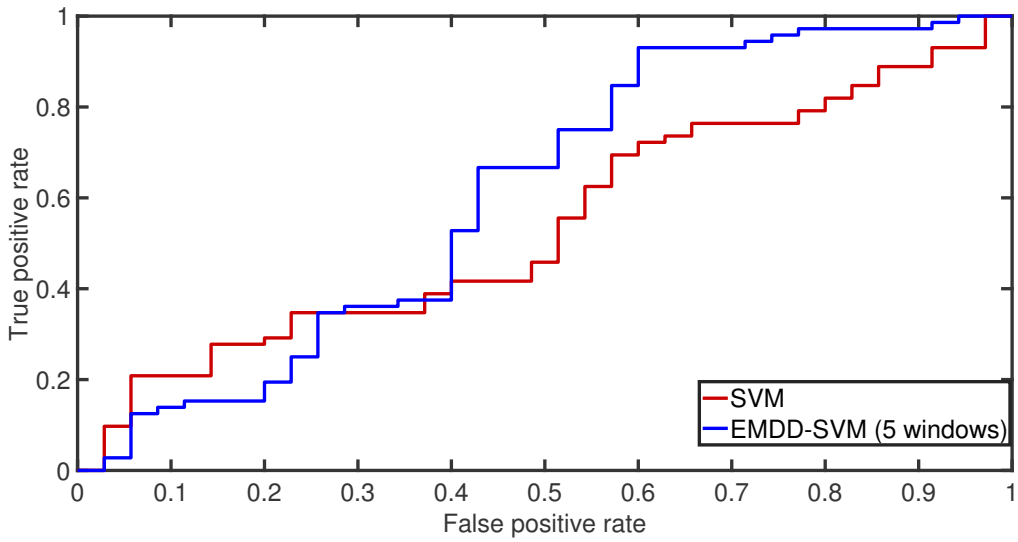


Figure 5.11.: ROC curve of the EMDD-SVM with $L = 5$ and the standard SVM approach over all participants for the *arousal* task

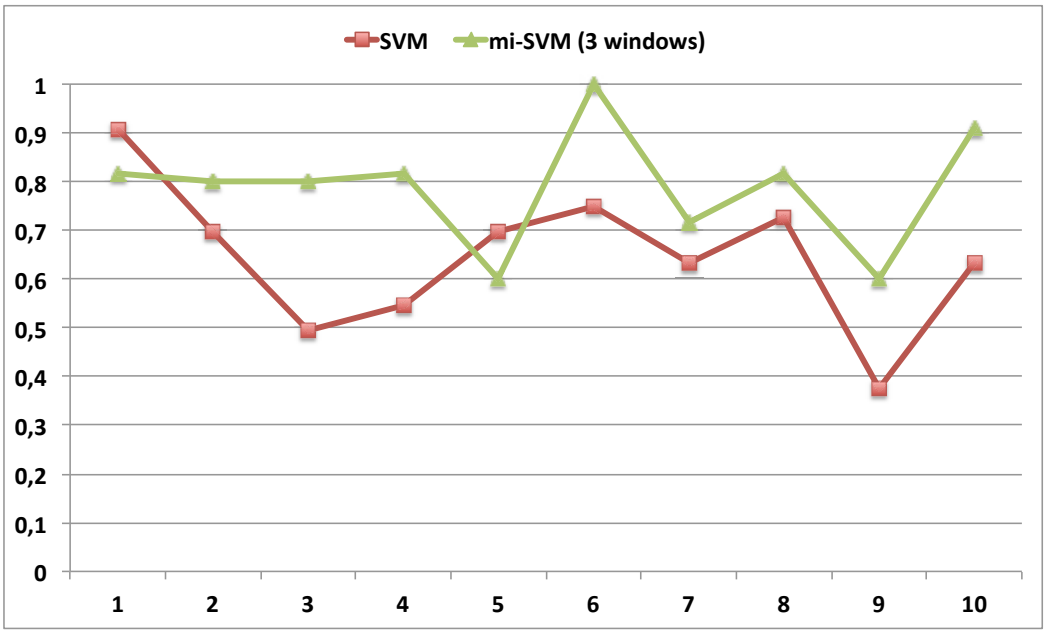


Figure 5.12.: The *macro-F1* score of the mi-SVM with 3 windows x video and the standard SVM approach for each participant for the *valence* task

Table 5.11.: Confusion matrices (rows are the true classes) of the mi-SVM with $L = 3$ and the standard SVM approach over all participants for the *valence* task

mi-SVM

	neg	*pos*
neg	0.84	0.16
pos	0.25	0.75

SVM

	neg	*pos*
neg	0.59	0.41
pos	0.29	0.71

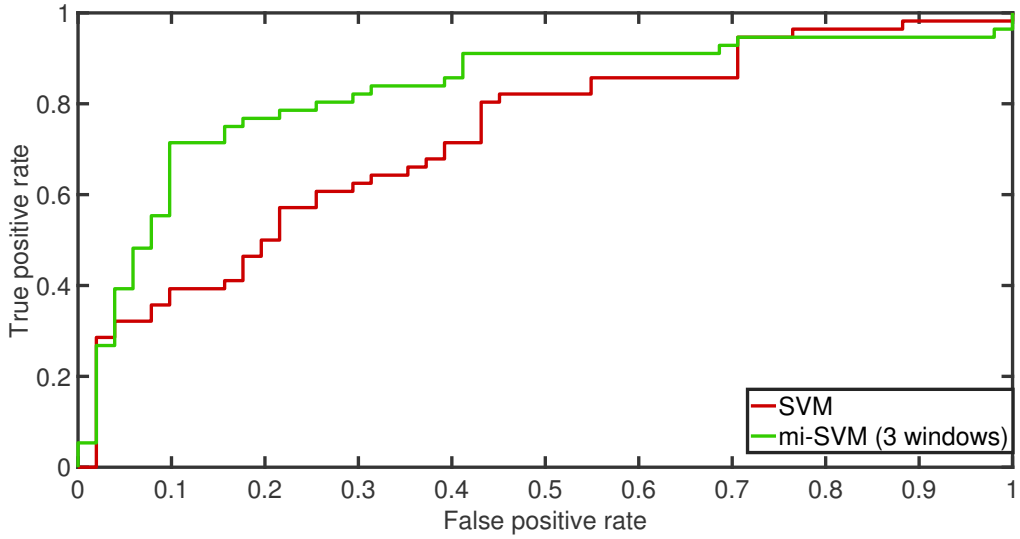


Figure 5.13.: ROC curve of the mi-SVM with $L = 3$ and the standard SVM approach over all participants for the *valence* task

Chapter 6.

Discussions

In this Section the author discusses and answers the research questions provided in Section 1.2.

6.1. Quantitative Assessment of Human Motion

The aim of this study was to provide an effective assessment tool to monitor and evaluate the correctness of rehabilitation exercises.

How can ML model be applied for assessing the human movement with respect to a reference example or a set of rules?

For this purpose, the proposed HSMM based approach allowed to combine different aspects of *template* and *rule* based methods, trying to overcome some of their limits. The goal is to obtain an approach clinically supported, reusable for a different set of exercises and computationally faster. As a *rule* based approach, the proposed algorithm is able to provide a quantitative score on the base of the main features defined by clinicians. At the same time, these features are evaluated with respect to an exemplar of the motion sequence, as a *template* based method. In particular, the exemplar is modeled by a probabilistic algorithm HSMM determined by the clinical judgment during the validation stage.

How can the algorithm provide a suitable feedback for supporting both clinicians and patients during rehabilitation process?

The features used for training the HSMM are based on modalities and purpose of the single exercise as indicated by expert clinicians. Hence, the proposed method can be considered as an interactive pattern recognition application that provides two main novelties: (1) it takes into account clinicians' indications (features) and (2) the amount of data to be analyzed, related only to those selected features, are lower than in algorithms of pattern recognition. These aspects make the data processing faster and clinically significant.

How can an objective measurement be designed to validate the proposed algorithm?

The reliability of the proposed approach is studied by evaluating its correlation with both clinical assessment and Dynamic Time Warping (DTW) algorithm, while healthy

and neurological disabled people performed a case study of five exercises included in a low-back pain physiotherapy program.

How far/close is the proposed algorithm from clinician evaluation of exercise performance?

The overall correlation obtained between HSMM and CS, proves that the presented approach is accurate for assessing human movement for different type of exercises. In particular, the medium-high correlation and the related low error values (i.e., RMS, MAE and MAPE) of those exercises (i.e., Exercises #1, #2) involving upper body, prove that this method could be extended to exercises for which the tracking sensor ensures better performances within its limits. This agrees with the well-known limits of Kinect system [260, 253, 254, 83] that lie in occlusion or overlaps of the joints, lower limb or no frontal plane tracking as it is possible to see in Exercises #3, #4, #5, where low correlation and higher error values have been obtained. Even if the selected features are the most significant movement descriptors, they can be affected by the subject position, presence of occlusions and environment conditions. In fact, among 41 enrolled subjects in the study, 2 subjects were completely lost for issues due to the Kinect sensor as reported in Figure 3.6. In detail, the results suggest that the skeleton tracking accuracy is higher for upper body compared to the lower one according to [253] and the accuracy validation study performed by author reported in Section 2.1.2.

How much better is the proposed algorithm with respect to standard algorithm widely used in literature?

In order to provide an index of the usability of the proposed HSMM approach on different types of exercise, the overall correlation (i.e., considering all the exercises scores), between the CS and HSMM scores, has been calculated $r = .60$, ($p < .01$). While, between CS and DTW, the overall correlation, $r = .56$ ($p < .01$), is lower.

For the ND group, the overall correlation of the HSMM score is higher (i.e., $r = .64$, $p < .01$) than DTW score (i.e., $r = 0.61$, $p < .01$). While for the HS is lower but significant (i.e., $r = .28$, $p = .01$) and for DTW is not significant (i.e., $r = .12$). Finally, between HSMM and DTW scores, the overall correlation is $r = .67$ ($p < .01$). The comparison between HSMM and DTW scores unveils a moderate significant correlation for HS (i.e., $r = .45$, $p =< .01$) and a high significant correlation for ND (i.e., $r = .7$, $p =< .01$). The results of the paper support that HSMM approach outperforms DTW, considered as one of the gold standard method for movement assessment [277, 74, 55]. A possible reason could be that the proposed approach avoids the problem of choosing a particular reference trajectory, modeling also temporally varying variance from the mean feature trajectory and reducing overfitting. Accordingly, the algorithm is able to quantitatively evaluate local erroneous movements. In particular, the local scores can be used to discriminate the accuracy of the gesture, providing specific feedback, useful both for clinicians and patients (see Figure 5.2).

Future direction may be addressed to go beyond the localization of the error (i.e.,

go beyond what the human/clinician sees) and provides an analysis of the local scores. This lead to extract the most salient feature for the assessment of movement for each different exercises. This analysis aims to:

- identify how the single feature and the related score contribute to the final outcome score.
- identify which feature was more discriminative between groups.

The canonical correlation analysis was able to weight the local score and then the extracted features well characterizing the total score. However, since author found the single scores are less powerful at discriminating between normal and pathological movement patterns, it is recommended that the sum of different local scores should be considered in order to make the proposed movement assessment a generalizable approach.

The comparison between groups shows how the method is able to better evaluate the performance of people affected by motor disabilities, while the correlation decreases remarkably when healthy subjects are considered. Moreover, HSMM provides results more correlated to CS than those obtained by DTW, for the ND group, highlighting the intrinsic variability connected to the patient as it is possible observing through SD and IQR results (see Table 5.4 and Figure 5.3). On the other hand, small errors, made by healthy subjects during the motion sequence, are more difficult to capture. Anyway, also the clinical judgments of the HS performance are less accurate and reproducible, because of difficulties to distinguish between small differences and possible anticipation biases.

6.1.1. Limits of the approach

The study presents some limits due to different aspects: first of all, the small sample size and the disease heterogeneity in the group of disabled people may limit the generalization power of the study. The trial is an explorative study, aimed at assessing the feasibility and reliability of the approach. All disabled subjects suffered from axial disorders, due to different pathologies, and the method resulted reliable in detecting posture difficulties during movement with respect to clinical judgment. The study is not powered to assess selective reliability for different nosographic conditions (diseases).

Secondarily the clinical judgment is provided through a questionnaire that never underwent a psychometric validation. For this reason, the author provided an inter-rater reliability of the questionnaire and analyzed its capacity to distinguish between healthy and disabled people. To the best of author's knowledge, in the literature, there is no questionnaire assessing the correctness of motor gesture during a training. This process is normally carried out by physiotherapists in real time during the training, looking at the patient performing movements. Normally, physiotherapists base the

judgment on their own experience following the exercise scope (as in a *rule* based manner).

In addition, the results provide large CI (95%) values, connected to RMSE, MAPE and MAE higher value, particularly for Exercises related to lower limb (see Exercises #4 and #5) or with overlaps of the joints (Exercise #3). This reduces the reliability of the assessment performed by data acquired through the Kinect v2, due to the intrinsic Microsoft Kinect Skeletal Tracking system limits, as above mentioned.

The system reliability has been tested on five exercises widely used to treat neurological and musculoskeletal diseases [265, 264]. The method followed in this study can be usefully applied for many other exercises (e.g., the University of Idaho-Physical Rehabilitation Movement Data [278] and the K3Da dataset [279]) adopted in the rehabilitation context: the procedure used to design the algorithm and to identify the outcome measures is the key rule to generalize the approach. Finally, the proposed approach, as for *template* based method, can be easily generalized and reused for a different set of rehabilitation exercises, once the salient features of the motor task to be assessed have been selected. Note that the features are exercise-specific and need to be defined according to the exercise scopes.

6.2. Emotion Inference from Physiological predictors

How can ML model be applied to infer the affective state of the user and model the variability in physiological response over the course of multimedia interaction? How can ML model be applied to handle the ambiguity and the change over the time of the emotional response?

The MIL approach assumes that the emotional response is ambiguous and summative over the entire multimedia interaction. This means that not all observation windows have the same predictive power. Then, author aims to find the time interval which leads to better prediction of the presence of the self-reported emotion. The computed results of two experimental databases (i.e., DEAP and Consumer database) and described in Table 5.6 and 5.9 confirm that a multi-instance based approach is appropriate for analyzing the physiological response in order to recognize human emotion. MIL methods provide the highest accuracy and *macro-F1* score for the *valence* task, respectively for the DEAP (average overall methods $\pm$ standard deviation: $ACC = 0.602 \pm 0.029$ and *macro-F1*=0.571 ± 0.034) and Consumer dataset ($ACC = 0.697 \pm 0.086$ and *macro-F1*=0.691 ± 0.088). The performed statistical analysis confirms the reliability of the proposed methodology for the estimation of positive and negative *valence* level: all MIL methods, except the EMDD-SVM with $L = 5$ shows a *macro-F1* significantly higher ($p < .05$) than chance level (.5) The *arousal* task shows a relatively lower accuracy and *macro-F1* score, according also to the re-

sults obtained by Koelstra et al. [138]. While in the Consumer dataset the unbalancing of class (16% low *arousal* vs 84% high *arousal*) can lead to a prediction bias and a classification of the majority class (see Table 5.10), in the DEAP scenario the predictive power of the features decreases and MIL methods sometimes fails to recognize the correct *arousal* state (see Figure 5.6). However, the mi-SVM and EMDD-SVM are able to predict respectively in the DEAP and Consumer dataset the *arousal* state with performance greater ($p < .05$) than chance level (.5). The estimation of the *dominance* in the DEAP context does not receive satisfactory results for both MIL and standard approaches. A deeper features extraction involving other more intrusive physiological signals (e.g., EEG, EMG) and video-content features extracted by the Multimedia Content Analysis (MCA, [280]) may be needed to solve this task.

Does ML method outperform standard supervised algorithm based on video-level features?

For what concern the *valence* task author has compared the best MIL approach with the related standard ML algorithm (i.e., SVM). For the DEAP dataset, the MI-SVM is superior than SVM method in terms of *macro-F1* score for 22/32 participants, *macro-F1* score (0.612 vs 0.557), accuracy (0.636 vs 0.581), AUC (0.660 vs 0.590), standard Precision (0.62 vs 0.57) and Recall (0.68vs0.64) (see Table 5.8). Accordingly, for the Consumer dataset, the mi-SVM shows greater performance than SVM (*macro-F1* score higher in $8/10$ fold, *macro-F1* score: 0.788 vs 0.646, accuracy: 0.789 vs 0.651, AUC: 0.829 vs 0.729, standard Precision (0.82 vs 0.63) and Recall (0.75 vs 0.71) (see Figure 5.11).

For what concern the *arousal* task, the standard SVM method is not significantly higher than chance level in the DEAP ($p = .076$) and Consumer dataset ($p = 0.957$). However, MIL methods such as mi-SVM with 3 windows x video and EMDD-SVM with 5 windows x video show a significant *macro-F1* distribution higher ($p < .05$) than chance level (.5) respectively for DEAP and Consumer dataset. In the DEAP dataset, the *macro-F1* score of mi-SVM (0.546) is higher than SVM (0.539) for 18/32 participants, as well as the accuracy (0.611 vs 0.591). Accordingly, the AUC (0.638 vs 0.636) and the standard Precision (0.57 vs 0.57) of the two approaches are comparable, while the Recall of mi-SVM is higher (0.71 vs 0.64) (see Table 5.7). For the Consumer dataset, the EMDD-SVM is superior to standard SVM in terms of *macro-F1* score for each participant (i.e., $7/10$ folds), *macro-F1* score (0.656 vs 0.497), accuracy (0.733 vs 0.629), AUC (0.610 vs 0.540), standard Precision (0.60 vs 0.51) and Recall (0.87 vs 0.83) (see Table 5.10).

Summarizing these outcomes, the proposed methodology based on EMDD-SVM, mi-SVM and MI-SVM approaches are reliable and accurate to estimate the binary *valence* state through the analysis of physiological signals. Furthermore, the *arousal* task is more difficult to solve and in this context, only some MIL methods (i.e., mi-SVM and EMDD-SVM) can solve it with an acceptable performance, still superior to standard supervised classifiers.

Is ML method reliable for the emotion recognition task towards the real world usage?

The obtained results from the Consumer dataset encourage the use of the proposed methodology towards the real world usage. Data were acquired by a commercial smartwatch and the experiment was performed in a less controlled environment (compared to that realized in Koelstra et al. [138]). In spite of the lower accuracy of the sensor and the less robustness of the labeling procedure, the proposed methodology is able to reliably predict the *valence* state and at the same time provides a reasonable estimation for the *arousal* state.

6.2.1. Label Assignment

The label assignment in the affective computing scenario is time-consuming and involves a relevant effort (see Section 2.2.4). Accordingly, a continuous labeling is needed in order to map the variability of the emotion during the multimedia interaction. The proposed application of Multiple Instance Learning for emotion recognition allows to overcome this drawback. In particular, the proposed MIL approach is able to model time interval which leads to better prediction of the presence of the self-reported emotion, without the needing of continuous labeling the subjective physiological response (as required by the sequential learning approaches). This allow to save time and effort in the label assignment procedure. Therefore, as future direction, the author aims to investigate the robustness of the presented MIL approach with respect to sequential supervised learning paradigm where the goal is to predict a new label sequence of emotion.

6.2.2. Relation to Number of Instances x Video

The MIL method is influenced by the size of the time interval: it is possible to model time interval which leads to better prediction of the presence of the self-reported emotion, capturing the subjective physiological response. For instance, for what concern the *arousal* task in the DEAP dataset, the mi-SVM with 3 and 5 windows x video shows different results: only the method with $L = 3$ provides a *macro-F1* score significantly higher than chance level. In the Consumer dataset different results are provided by the two configurations of EMDD-SVM for the *arousal* task. In this case, the method with $L = 5$ provides a *macro-F1* score significantly higher than chance level while the $L = 3$ setting provides a lower and not significant *macro-F1*. Finally, the choice of the number of instances for each video can change for each subject and depend on several factors such as the type and the duration of stimuli and when the physiological response occurs. In the future, the author would like to select the correct number of instances maximizing information theory metrics such as entropy and mutual information.

Chapter 7.

Conclusions

The main contribution of this thesis is the application of different machine learning algorithms for monitoring the physical and emotional well-being of a subject. Two problems were formulated and answered:

1. quantitative assessment of human motion: real-time evaluation of the exercise performance during physical rehabilitation stage;

2. emotion inference from physiological predictors: inferring the affective state of the user during multimedia interaction.

The HSMM based approach is proposed for an accurate rehabilitation exercise assessment, close to the clinical evaluation. The HSMM algorithm resulted in a viable solution to provide an effective and reliable quantitative feedback to physiotherapists and patients. Sensor characteristics limit the performance of the approach in exercises with movements on the deep spatial plane (i.e., Exercise #4) or with joints overlapping in the sensor's view (i.e., Exercise #3) or related to the low limb (i.e., Exercise #5). From a clinical point of view, the introduction of a detailed posture control beyond the assessment of the gesture goal achievement is a novelty, useful especially in the rehabilitation of the axial control and of the low-back pain.

Additionally, the proposed HSMM method

- outperforms DTW based methods presented in [55] and [92], demonstrating its applicability in movement assessment;

- allows to provide a score in few seconds that can be useful for telerehabilitation;

- is able to distinguish between patients and healthy subjects;

- allow the clinician to localize the error in the exercise movement execution;

- can be generalized to different types of exercises.

To achieve a practical impact, the proposed algorithm could be embedded in a rehabilitation framework, for a remote assessment, composed by a TV, a computer, and

a network connection. The score could then be used to provide immediate and highly specific feedback to patients and medical centers in a cloud network.

The MIL model-based approaches are proposed to improve the prediction of emotional state during multimedia interaction using the physiological signals as inputs. In particular, typical approaches do not consider that emotions events are often momentarily within a given time-window rather than pervasive though it. The proposed MIL approach aims to improve classification by taking into account this specific aspect. This is very critical in real life applications where labeling of data is sparse and possibly describing only the more important event rather than the typical continuous subtle affective changes that occurs. Two databases of physiological signals were considered to test the MIL model: the former is the DEAP database, a gold standard database collected in [138], while the latter is a dataset collected by the author closer to real-world condition and problematic. The obtained results point out the reliability of the proposed methodology in the gold standard scenario and in the real-world usage. Future works could be addressed to consider also an *user-independent* MIL models able to generalize across different users. Since it is difficult to obtain remarkable results with an *user-independent* approach additional features are needed. A viable solution may be found combining physiological features with audio-video features in a multi-view learning scenario. Then, the combination of Multi-Instance and Multi Kernel techniques could at the same time: (i) localize the physiological response identifying the time segment which leads to better prediction of the presence of the judged emotion and (ii) automatically weight the importance of each different feature, in order to improve the algorithm performance.

Future works may be addressed to localize the time interval of the video stimuli in which the physiological pattern of interest is more strongly displayed. This leads to not only model and discriminate the self-reported emotion but also to provide its localization. Additionally, we aim to evaluate the proposed methodology on different affective recognition topics. For instance, the automatic pain detection from a patients'face expression represents one such application, and the UNBC-McMaster Shoulder Pain Expression Archive Database [281] can be evaluated. Another interesting future direction is to extend the methodology into multi-instance multi-label formulation [282] where the emotional response is described by multiple instances and associated with multiple class labels. Accordingly, the MIL approach could be extended to map the dimensional perspectives of the emotion. The natural extension is to formulate the problem as a multiple instance regression task [283, 284, 285].

Finally, as future works, the emotional effects on movement execution in a patient with chronic pain could be investigated. Then, the two proposed approaches for motion evaluation and emotion recognition could be combined in order to provide a biomechanical score and an evaluation of pain-related affective emotions, following the procedure reported in Section 2.3. This could be relevant for the segmentation of

the movement too, as people may hesitate to perform the movement due to fear of injury or lack of confidence in movement capabilities.

Appendix A.

Questionnaire: Exercise accuracy assessment

Please, observing the entire exercise (all repetitions), answer the questions signing one of the following chance:

1=Never	2=Rarely	3=Sometimes	4=Often	5=Always

1. Is the primary goal of the exercise reached (i.e., the extension of the upper limbs, trunk rotation with upper limbs elevated to 90°, squatting, etc.)?
2. Is the exercise repeatable?
3. Is the amplitude of the movement complete?
4. Is the posture of the head correct?
5. Is the posture of the right arm correct?
6. Is the posture of the left arm correct?
7. Is the posture of the trunk correct?
8. Is the posture of the pelvis correct?
9. Is the posture of the right leg correct?
10. Is the posture of the left leg correct?

Appendix B.

Pseudocode MIL methods

Algorithm 1 EMDD-SVM

Main(Bag)
  let $k = 10$ global // number of different starting Bags
  let $C = [0.1, 0.5, 1, 5, 25, 100]$ global; // SVM Box Constraint
  define *num_inst* as number of instances for each bag
  $[B, B\_test]$=**CV_partition**$(Bag, nfold1)$
  for $i = 1$, $i{+}{+}$, while $i < nfold1$
    $[Model, h]$=**Training**(B, y)
    $[acc(i), F_1(i), CM\{i\}]$=**Test**$(B\_test, y\_test, Model, h)$
end

$[Model, h]$=**Training**(B, y, k)
  pick k random positive bags $B_1, \ldots, B_k$ from B
  for $q = 1$, $q{+}{+}$, while $q < k * num\_inst$
    [c(q), s(q), DD(q)]=EMDD(BagTrainIn,k) // EMDD [223]
  $loc\_DD\_max = \arg\max_q(DD(q))$
  $h = [c(loc\_max), s(loc\_max)]$
  compute $\phi(B^-)$ and $\phi(B^+)$
  (B_in, B_val)=**CV_partition**$(B, nfold2)$ // Validation stage
  for $r = 1$, $r{+}{+}$, while $r < nfold2$
    compute $P(h|B\_in^-)$ and $P(h|B\_in^+)$
    compute $\phi(B\_in^-)$ and $\phi(B\_in^+)$
    for $s = 1$, $s{+}{+}$, while $s < length(C)$
      $Model\_val$ = fitcsvm$(\phi(B_in^-), \phi(B_in^+), C(s))$
      $F_1\_val(r,s)$=**Test**(B_val, y_val, Model_val)
  $F_1\_val\_avg$=mean$_r(F_1_val)$
  $loc\_C\_max = \arg\max_s(F_1\_val\_avg)$
  $C\_max = C(loc\_C\_max)$
  $Model$ = fitcsvm$(\phi(B^-), \phi(B^+), C_max)$

Algorithm 1 EMDD-SVM

$[acc, F_1, CM]$=**Test**$(B\_test, y\_test, Model, h)$
  compute $P(h|B\_test)$
  compute $\phi(B\_test)$
  compute $score_i = \boldsymbol{w}^T \phi(B\_test_i) + b$ for each $\phi(B\_test_i)$
  if $score_i > 0$
    $pred\_y\_test_i = 1$
  else
    $pred\_y\_test_i = -1$
  compute acc, F_1, CM

$[c(q), s(q), DD(q)] = EMDD(B, k)$
  let $h = [c, s]$ // initial hypothesis
  set $nldd_0 = \infty$
  set $nldd_1 = NLDD(h, B)$
  repeat
    compute $p_i^* = \arg\max_j (P(h|B_{ij})$ for each B_{ij} // E step
    $h' = \arg\max_h \left[\prod_{i=1}^{l} P(c|p_i^{*+}) \prod_{i=1}^{m} P(c|p_i^{*-}) \right]$ // M step
    $nldd_0 = nldd_1$
    $nldd_1 = NLDD(h', B)$
    $h = h'$
  while $nldd_1 < nldd_0$

Algorithm 2 mi-SVM

Main(Bag)
  let $C = [0.1, 0.5, 1, 5, 25, 100]$ global; // SVM Box Constraint
  let $th = [-2, -1.999, \ldots, 2]$ global; // SVM score threshold
  $[B, B\_test]$=**CV_partition**$(Bag, nfold1)$
  for $i = 1$, i++, while $i < nfold1$
    $[Model, th\_max]$=**Training**(B, y)
    $[acc(i), F_1(i), CM\{i\}]$=**Test**$(B\_test, y\_test, Model, th\_max)$
end

$[Model, th\_max$=**Training**(B, y)
  (B_in, B_val)=**CV_partition**$(B, nfold2)$ // Validation stage
  for $r = 1$, r++, while $r < nfold2$
    for $s = 1$, s++, while $s < length(C)$
    $Model\_val$ = mi-SVM$(B_in, y_in, C(s))$; // mi-SVM [215]
    for $t = 1$, t++, while $t < length(threshold)$
      $F_1\_val(r,s,t)$=**Test**$(B\_val, y\_val, Model\_val, th(t))$
  $F_1\_val\_avg$=mean$_r(F_1_val(r,s,t))$
  $[loc\_C\_max, loc\_th\_max] = \arg\max_{s,t}(F_1\_val\_avg)$
  $C\_max = C(loc\_C\_max)$
  $th\_max = th(loc\_th\_max)$
  $Model$ = mi-SVM(B, y, C_max); // mi-SVM [215]

$[acc, F_1, CM]$=**Test**$(B\_test, y\_test, Model, th)$
  compute $score_{ij}$ for each x_test_{ij}
  $score_i = \max_j(score_{ij})$
  if $score_i > th$
    $pred\_y\_test_i = 1$
  else
    $pred\_y\_test_i = -1$
  compute acc, F_1, CM

$Model$ = mi-SVM(B, y, C_max)
  repeat
    $Model$=SVM(B, y, C_max);
    compute Model $score_{ij}$ for each B_{ij}
    for each B^+
      if $(\sum_j(1+y_{ij})/2 == 0)$
        compute $j^* = \arg\max_j(score_{ij})$
        set $y_{j^*} = 1$
  while imputed labels have changed

Algorithm 3 MI-SVM

Main(Bag)
  let $C = [0.1, 0.5, 1, 5, 25, 100]$ global; // SVM Box Constraint
  let $th = [-2, -1.999, \ldots, 2]$ global; // SVM score threshold
  $[B, B\_test]$=**CV_partition**$(Bag, nfold1)$
  for $i = 1$, $i{+}{+}$, while $i < nfold1$
    $[Model, th\_max]$=**Training**(B, y)
    $[acc(i), F_1(i), CM\{i\}]$=**Test**$(B\_test, y\_test, Model, th\_max)$
end

$[Model, th\_max]$=**Training**(B, y)
  (B_in, B_val)=**CV_partition**$(B, nfold2)$ // Validation stage
  for $r = 1$, $r{+}{+}$, while $r < nfold2$
    for $s = 1$, $s{+}{+}$, while $s < length(C)$
    $Model\_val$ = MI-SVM$(B_in, y_in, C(s))$; // MI-SVM [215]
    for $t = 1$, $t{+}{+}$, while $t < length(threshold)$
      $F_1\_val(r,s,t)$=**Test**$(B\_val, y\_val, Model\_val, th(t))$
  $F_1\_val\_avg$=mean$(F_1_val(r,s,t))$
  $[loc\_C\_max, loc\_th\_max] = \arg\max_{s,t}(F_1\_val\_avg)$
  $C\_max = C(loc\_C\_max)$
  $th\_max = th(loc\_th\_max)$
  $Model$ = MI-SVM(B, y, C_max); // MI-SVM [215]

$[acc, F_1, CM]$=**Test**$(B\_test, y\_test, Model, th)$
  compute $score_{ij}$ for each B_test_{ij}
  $score_i = \max_j(score_{ij})$
  if $score_i > th$
    $pred\_y\_test_i = 1$
  else
    $pred\_y\_test_i = -1$
  compute acc, F_1, CM

$Model$ = MI-SVM(B, y, C_max)
  set $B_i^+ = \frac{\sum_j x_{ij}^+}{|B_i^+|}$
  repeat
    B_new=$[x_{ij}^-; B_i^+]$;
    $Model$=SVM(B_new, y, C_max);
    compute Model $score_{ij}$ for each x_{ij}^+
    compute $s(i) = \arg\max_j(score_{ij})$ for each B_i^+
    set $B_i^+ = x_{is(i)}$
  while selector variables have changed

Bibliography

[1] J. Stiglitz, A. Sen, J.-P. Fitoussi *et al.*, "The measurement of economic performance and social progress revisited," *Sciences Po publications*, 2009.

[2] WHO, "World health statistics 2015," World Health Organization, Tech. Rep., 2015.

[3] S. Stewart-Brown, "Emotional wellbeing and its relation to health: Physical disease may well result from emotional distress," *BMJ: British Medical Journal*, vol. 317, no. 7173, p. 1608, 1998.

[4] I. G. I. for Ethical Considerations in Artificial Intelligence and A. Systems, "Prioritizing of human well-being in the age of artificial intelligence," IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems, Tech. Rep., 2017.

[5] T. R. Society, "Machine learning: the power and promise of computers that learn by example," The Royal Society, Tech. Rep., 2017.

[6] S. Huang, J. Zhou, Z. Wang, Q. Ling, and Y. Shen, "Biomedical informatics with optimization and machine learning," *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2017, no. 1, p. 4, Feb 2017.

[7] WHO, "Health in 2015: from MDGs to SDGs," World Health Organization, Tech. Rep., 2015.

[8] H. Zhou and H. Hu, "Human motion tracking for rehabilitation—a survey," *Biomedical Signal Processing and Control*, vol. 3, no. 1, p. pp. 1–18, 2008.

[9] S. Patel, H. Park, P. Bonato, L. Chan, and M. Rodgers, "A review of wearable sensors and systems with application in rehabilitation," *Journal of neuroengineering and rehabilitation*, vol. 9, no. 1, p. 21, 2012.

[10] J. M. Winters, Y. Wang, and J. M. Winters, "Wearable sensors and telerehabilitation," *IEEE Engineering in Medicine and Biology Magazine*, vol. 22, no. 3, pp. 56–65, May 2003.

[11] H. Yan, H. Huo, Y. Xu, and M. Gidlund, "Wireless sensor network based e-health system: Implementation and experimental results," *IEEE Transactions on Consumer Electronics*, vol. 56, no. 4, pp. 2288–2295, November 2010.

[12] J. Wang, Z. Zhang, B. Li, S. Lee, and R. S. Sherratt, "An enhanced fall detection system for elderly person monitoring using consumer home networks," *IEEE Transactions on Consumer Electronics*, vol. 60, no. 1, pp. 23–29, February 2014.

[13] P. Afsar, P. Cortez, and H. Santos, "Automatic visual detection of human behavior: A review from 2000 to 2014," *Expert Systems with Applications*, vol. 42, no. 20, pp. 6935 – 6956, 2015.

[14] S. Elloumi, S. Cosar, G. Pusiol, F. Bremond, and M. Thonnat, "Unsupervised discovery of human activities from long-time videos," *IET Computer Vision*, vol. 9, no. 4, pp. 522–530, 2015.

[15] Z. A. Khan and W. Sohn, "Abnormal human activity recognition system based on r-transform and kernel discriminant technique for elderly home care," *IEEE Transactions on Consumer Electronics*, vol. 57, no. 4, pp. 1843–1850, November 2011.

[16] A. Jalal, M. Z. Uddin, and T. S. Kim, "Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 3, pp. 863–871, August 2012.

[17] S. Gaglio, G. L. Re, and M. Morana, "Human activity recognition process using 3-d posture data," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 5, pp. 586–597, Oct 2015.

[18] S. Chun and C. S. Lee, "Human action recognition using histogram of motion intensity and direction from multiple views," *IET Computer Vision*, vol. 10, no. 4, pp. 250–256, 2016.

[19] M. Selmi, M. A. El-Yacoubi, and B. Dorizzi, "Two-layer discriminative model for human activity recognition," *IET Computer Vision*, vol. 10, no. 4, pp. 273–278, 2016.

[20] Y. Yi, Z. Zheng, and M. Lin, "Realistic action recognition with salient foreground trajectories," *Expert Systems with Applications*, pp. –, 2017.

[21] S. M. Yoon and A. Kuijper, "Human action recognition based on skeleton splitting," *Expert Systems with Applications*, vol. 40, no. 17, pp. 6848 – 6855, 2013.

[22] A. A. Chaaraoui, J. R. Padilla-López, P. Climent-Pérez, and F. Flórez-Revuelta, "Evolutionary joint selection to improve human action recognition with rgb-d devices," *Expert Systems with Applications*, vol. 41, no. 3, pp. 786 – 794, 2014.

[23] M. Faraki, M. Palhang, and C. Sanderson, "Log-euclidean bag of words for human action recognition," *IET Computer Vision*, vol. 9, no. 3, pp. 331–339, 2015.

[24] S. Spinsante and E. Gambi, "Remote health monitoring by osgi technology and digital tv integration," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 4, pp. 1434–1441, November 2012.

[25] C. H. Hung, Y. W. Bai, and R. Y. Tsai, "Design of blood pressure measurement with a health management system for the aged," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 2, pp. 619–625, May 2012.

[26] K. Cranen, G. C. Groothuis-Oudshoorn, M. M. Vollenbroek-Hutten, and J. M. IJzerman, "Toward patient-centered telerehabilitation design: Understanding chronic pain patients' preferences for web-based exercise telerehabilitation using a discrete choice experiment," *J Med Internet Res*, vol. 19, no. 1, p. e26, Jan 2017.

[27] J. F. S. Lin and D. Kulić, "Online segmentation of human motion for automated rehabilitation exercise analysis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 1, 2014.

[28] P. Fernandez de Dios, P. Chung, and Q. Meng, "Landmark-based methods for temporal alignment of human motions," *IEEE Computational Intelligence Magazine*, vol. 9, no. 2, pp. 29–37, May 2014.

[29] R. Vertegaal *et al.*, "Attentive user interfaces," *Communications of the ACM*, vol. 46, no. 3, pp. 30–33, 2003.

[30] P. H. Blaney, "Affect and memory: a review." *Psychological bulletin*, vol. 99, no. 2, p. 229, 1986.

[31] G. H. Bower, S. G. Gilligan, and K. P. Monteiro, "Selectivity of learning caused by affective states." *Journal of Experimental Psychology: General*, vol. 110, no. 4, p. 451, 1981.

[32] B. Reeves and C. Nass, *How people treat computers, television, and new media like real people and places*. CSLI Publications and Cambridge university press, 1996.

[33] M. Lewis and J. Haviland-Jones, *Handbook of Emotions*. Guilford Publications, 2000. [Online]. Available: https://books.google.it/books?id=mxDzOSvivIkC

[34] A. Bechara, "The role of emotion in decision-making: Evidence from neurological patients with orbitofrontal damage," *Brain and Cognition*, vol. 55, no. 1, pp. 30 – 40, 2004, development of Orbitofrontal Function.

[35] N. Schwarz, *Emotion, Cognition, and Decision Making*, ser. Cognition & emotion. Psychology Press, 2000. [Online]. Available: https://books.google.it/books?id=NLTHGwAACAAJ

[36] J. A. Jacko, *Human computer interaction handbook: Fundamentals, evolving technologies, and emerging applications*. CRC press, 2012.

[37] H. Dunbar and J. F. Josiah Macy, *Emotions and Bodily Changes: A Survey of Literature on Psychosomatic Interrelationships, 1910-1933*. Josiah Macy, Jr., Foundation, 1938. [Online]. Available: https://books.google.it/books?id=HB60AAAAIAAJ

[38] N. H. Eller, J. Kristiansen, and Å. M. Hansen, "Long-term effects of psychosocial factors of home and work on biomarkers of stress," *International Journal of Psychophysiology*, vol. 79, no. 2, pp. 195–202, 2011.

[39] N. Jackman, W. Schottstaedt, S. C. McPhal, and S. Wolf, "Interaction, emotion and physiologic change," *Journal of health and human behavior*, pp. 83–87, 1963.

[40] A. Kaklauskas, E. K. Zavadskas, V. Pruskus, A. Vlasenko, L. Bartkiene, R. Paliskiene, L. Zemeckyte, V. Gerstein, G. Dzemyda, and G. Tamulevicius, "Recommended biometric stress management system," *Expert Systems with Applications*, vol. 38, no. 11, pp. 14 011–14 025, 2011.

[41] "American institute of stress," http://www.stress.org/daily-life/.

[42] OSHwiki, "Rehabilitation and return-to-work polices and systems in european countries — oshwiki,," 2016, [Online; accessed 8-August-2016]. [Online]. Available: http://oshwiki.eu/index.php?title=Rehabilitation_and_return-to-work_polices_and_systems_in_European_Countries&oldid=246065

[43] "Atlas of emotions." [Online]. Available: http://atlasofemotions.org

[44] U. Neisser, "The imitation of man by machine," *Science*, vol. 139, no. 3551, pp. 193–197, 1963.

[45] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[46] "Affective computing csci534, jonathan gratch," http://people.ict.usc.edu/~gratch/CSCI534/CSCI534-Syllabus.pdf, accessed: 2017-11-07.

[47] P. Daponte, L. De Vito, M. Riccio, and C. Sementa, "Design and validation of a motion-tracking system for rom measurements in home rehabilitation," *Measurement*, vol. 55, pp. 82–96, 2014.

[48] P. Arpaia, P. Cimmino, E. De Matteis, and G. D'Addio, "A low-cost force sensor-based posturographic plate for home care telerehabilitation exergaming," *Measurement*, vol. 51, pp. 400–410, 2014.

[49] M. van Diest, C. Lamoth, J. Stegenga, G. Verkerke, and K. Postema, "Exergaming for balance training of elderly: state of the art and future developments," *Journal of NeuroEngineering and Rehabilitation*, vol. 10, no. 1, p. 101, 2013.

[50] M. van Diest, J. Stegenga, H. J. Wörtche, K. Postema, G. J. Verkerke, and C. J. Lamoth, "Suitability of kinect for measuring whole body movement patterns during exergaming," *Journal of Biomechanics*, vol. 47, no. 12, pp. 2925 – 2932, 2014.

[51] M. Kutlu, C. Freeman, E. Hallewell, A.-M. Hughes, and D. Laila, "Upper-limb stroke rehabilitation using electrode-array based functional electrical stimulation with sensing and control innovations," *Medical Engineering & Physics*, vol. 38, no. 4, pp. 366 – 379, 2016.

[52] C. Metcalf, R. Robinson, A. Malpass, T. Bogle, T. Dell, C. Harris, and S. Demain, "Markerless motion capture and measurement of hand kinematics: Validation and application to home-based upper limb rehabilitation," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 8, pp. 2184–2192, Aug 2013.

[53] G. Palacios-Navarro, I. García-Magariño, and P. Ramos-Lorente, "A kinect-based system for lower limb rehabilitation in parkinson's disease patients: a pilot study," *Journal of Medical Systems*, vol. 39, no. 9, pp. 1–10, 2015.

[54] Y.-J. Chang, W.-Y. Han, and Y.-C. Tsai, "A kinect-based upper limb rehabilitation system to assist people with cerebral palsy," *Research in Developmental Disabilities*, vol. 34, no. 11, pp. 3654 – 3659, 2013.

[55] C.-J. Su, C.-Y. Chiang, and J.-Y. Huang, "Kinect-enabled home-based rehabilitation system using dynamic time warping and fuzzy logic," *Applied Soft Computing*, vol. 22, pp. 652 – 666, 2014.

[56] D. González-Ortega, F. Díaz-Pernas, M. Martínez-Zarzuela, and M. Antón-Rodríguez, "A kinect-based system for cognitive rehabilitation exercises monitoring," *Computer Methods and Programs in Biomedicine*, vol. 113, no. 2, pp. 620 – 631, 2014.

[57] R. Kizony, P. L. Weiss, S. Harel, Y. Feldman, A. Obuhov, G. Zeilig, and M. Shani, "Telerehabilitation service delivery journey from prototype to robust in-home use," *Disability and rehabilitation*, vol. 39, no. 15, pp. 1532–1540, 2017.

[58] B. Lange, S. Koenig, C.-Y. Chang, E. McConnell, E. Suma, M. Bolas, and A. Rizzo, "Designing informed game-based rehabilitation tasks leveraging advances in virtual reality," *Disability and rehabilitation*, vol. 34, no. 22, pp. 1863–1870, 2012.

[59] T. G. Russell, P. Buttrum, R. Wootton, and G. A. Jull, "Internet-based outpatient telerehabilitation for patients following total knee arthroplasty," *The Journal of Bone & Joint Surgery*, vol. 93, no. 2, pp. 113–120, 2011.

[60] V. Venkatesh, M. G. Morris, G. B. Davis, and F. D. Davis, "User acceptance of information technology: Toward a unified view," *MIS Quarterly*, vol. 27, no. 3, pp. 425–478, 2003.

[61] M. Iosa, P. Picerno, S. Paolucci, and G. Morone, "Wearable inertial sensors for human movement analysis," *Expert Review of Medical Devices*, vol. 13, no. 7, pp. 641–659, 2016.

[62] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1995–2006, 2013.

[63] J. Shotton and et al., "Efficient human pose estimation from single depth images," *Trans. PAMI*, 2012.

[64] J. Sell and P. O'Connor, "The xbox one system on a chip and kinect sensor," *IEEE Micro*, vol. 34, no. 2, 2014.

[65] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3d human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 914–927, May 2014.

[66] H.-W. Lee, C.-H. Liu, K.-T. Chu, Y.-C. Mai, P.-C. Hsieh, K.-C. Hsu, and H.-C. Tseng, "Kinect who's coming—applying kinect to human body height measurement to improve character recognition performance," *Smart Science*, vol. 3, no. 2, pp. 117–121, 2015.

[67] E. E. Stone and M. Skubic, "Fall detection in homes of older adults using the microsoft kinect," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 290–301, Jan 2015.

[68] S. Erik and S. Marjorie, "Unobtrusive, continuous, in-home gait measurement using the microsoft kinect," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 10, pp. 2925–2932, 2013.

[69] C. Morrison, P. Culmer, H. Mentis, and T. Pincus, "Vision-based body tracking: turning kinect into a clinical tool," *Disability and Rehabilitation: Assistive Technology*, vol. 11, no. 6, pp. 516–520, 2016.

[70] B. Seamon, M. DeFranco, and M. Thigpen, "Use of the xbox kinect virtual gaming system to improve gait, postural control and cognitive awareness in an individual with progressive supranuclear palsy," *Disability and Rehabilitation*, pp. 1–6, 2016.

[71] "Skeletal joint smoothing white paper kinect for windows 1.5, 1.6, 1.7, 1.8," https://msdn.microsoft.com/en-us/library/jj131429.aspx, accessed: 2017-09-21.

[72] J. Shu, F. Hamano, and J. Angus, "Application of extended kalman filter for improving the accuracy and smoothness of kinect skeleton-joint estimates," *Journal of Engineering Mathematics*, vol. 88, no. 1, pp. 161–175, 2014.

[73] W. Zhao, R. Lun, D. D. Espy, and M. A. Reinthal, "Rule based realtime motion assessment for rehabilitation exercises," in *Computational Intelligence in Healthcare and e-health (CICARE), 2014 IEEE Symposium on*, Dec 2014, pp. 133–140.

[74] M. C. Hu, C. W. Chen, W. H. Cheng, C. H. Chang, J. H. Lai, and J. L. Wu, "Real-time human movement retrieval and assessment with kinect sensor," *IEEE Transactions on Cybernetics*, vol. 45, no. 4, pp. pp. 742–753, 2015.

[75] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 3, pp. pp. 412–424, March 2012.

[76] W. Ding, K. Liu, X. Fu, and F. Cheng, "Profile hmms for skeleton-based human action recognition," *Signal Processing: Image Communication*, vol. 42, pp. pp. 109 – 119, 2016.

[77] R. A. Clark, Y.-H. Pua, K. Fortin, C. Ritchie, K. E. Webster, L. Denehy, and A. L. Bryant, "Validity of the microsoft kinect for assessment of postural control," *Gait & Posture*, vol. 36, no. 3, pp. 372 – 377, 2012.

[78] R. A. Clark, Y.-H. Pua, A. L. Bryant, and M. A. Hunt, "Validity of the microsoft kinect for providing lateral trunk lean feedback during gait retraining," *Gait & posture*, vol. 38, no. 4, pp. 1064–1066, 2013.

[79] R. A. Clark, K. J. Bower, B. F. Mentiplay, K. Paterson, and Y.-H. Pua, "Concurrent validity of the microsoft kinect for assessment of spatiotemporal gait variables," *Journal of biomechanics*, vol. 46, no. 15, pp. 2722–2725, 2013.

[80] A. P. L. Bo, M. Hayashibe, and P. Poignet, "Joint angle estimation in rehabilitation with inertial sensors and its integration with kinect," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE, 2011, pp. 3479–3483.

[81] E. Akdoğan, E. Taçgın, and M. A. Adli, "Knee rehabilitation using an intelligent robotic system," *Journal of Intelligent Manufacturing*, vol. 20, no. 2, p. 195, Jan 2009.

[82] Q. Wang, P. Turaga, G. Coleman, and T. Ingalls, "Somatech: an exploratory interface for altering movement habits," in *Proceedings of the extended abstracts of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 2014, pp. 1765–1770.

[83] A. K. Mishra, M. Skubic, and C. Abbott, "Development and preliminary validation of an interactive remote physical therapy system," in *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug 2015, pp. 190–193.

[84] B. C. Bedregal, A. C. Costa, and G. P. Dimuro, "Fuzzy rule-based hand gesture recognition," in *IFIP International Conference on Artificial Intelligence in Theory and Practice*. Springer, 2006, pp. 285–294.

[85] T. Hachaj and M. R. Ogiela, "Rule-based approach to recognizing human body poses and gestures in real time," *Multimedia Systems*, vol. 20, no. 1, pp. 81–99, 2014.

[86] M. Capecci, M. Ceravolo, F. D'Orazio, F. Ferracuti, S. Iarlori, G. Lazzaro, S. Longhi, L. Romeo, and F. Verdini, "A tool for home-based rehabilitation allowing for clinical evaluation in a visual markerless scenario," in *EMBC, 37th Annual Int. Conf. of the IEEE*, Aug 2015.

[87] L. Ciabattoni, F. Ferracuti, S. Iarlori, S. Longhi, and L. Romeo, "A novel computer vision based e-rehabilitation system: From gaming to therapy support," in *IEEE Int. Conf. on Consumer Electronics (ICCE)*, Jan 2016, pp. 43–44.

[88] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, Nov 2008.

[89] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. pp. 976 – 990, 2010.

[90] S. Nomm and K. Buhhalko, "Monitoring of the human motor functions rehabilitation by neural networks based system with kinect sensor," in *IFAC Proceedings Volumes (IFAC-PapersOnline)*, vol. 46, no. 15, 2013, pp. 249–253.

[91] J. F. S. Lin and D. Kulić, "Online segmentation of human motion for automated rehabilitation exercise analysis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 1, pp. pp. 168–180, Jan 2014.

[92] Z. Zhang, Q. Fang, and X. Gu, "Objective assessment of upper-limb mobility for poststroke rehabilitation," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 4, pp. pp. 859–868, April 2016.

[93] O. Celiktutan, C. B. Akgul, C. Wolf, and B. Sankur, "Graph-based analysis of physical exercise actions," in *Proceedings of the 1st ACM International Workshop on Multimedia Indexing and Information Retrieval for Healthcare*, ser. MIIRH '13. New York, NY, USA: ACM, 2013, pp. 23–32.

[94] R. Wang, G. Medioni, C. J. Winstein, and C. Blanco, "Home monitoring musculo-skeletal disorders with a single 3d sensor," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2013.

[95] M. Nieto-Hidalgo, F. J. Ferrández-Pastor, R. J. Valdivieso-Sarabia, J. Mora-Pascual, and J. M. García-Chamizo, "A vision based proposal for classification of normal and abnormal gait using rgb camera," *Journal of Biomedical Informatics*, vol. 63, pp. pp. 82 – 89, 2016.

[96] D. Leightley, M. H. Yap, and J. McPhee, "Automated analysis and quantification of human mobility using a depth sensor," *IEEE Journal of Biomedical and Health Informatics*, vol. PP, no. 99, 2016.

[97] I. Pernek, G. Kurillo, G. Stiglic, and R. Bajcsy, "Recognizing the intensity of strength training exercises with wearable sensors," *Journal of Biomedical Informatics*, vol. 58, pp. pp. 145 – 155, 2015.

[98] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. Tygar, "Adversarial machine learning," in *Proceedings of the 4th ACM workshop on Security and artificial intelligence*. ACM, 2011, pp. 43–58.

[99] G. H. John, "Robust decision trees: Removing outliers from databases." in *KDD*, 1995, pp. 174–179.

[100] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artificial intelligence review*, vol. 22, no. 2, pp. 85–126, 2004.

[101] V. Zatsiorsky, *Kinetics of Human Motion*. Human Kinetics, 2002, last accessed: 2017/02/23. [Online]. Available: https://books.google.it/books?id=wp3zt7oF8a0C

[102] D. B. Shalom, S. H. Mostofsky, R. L. Hazlett, and M. C. G. R. J. L. Y. F. R. Hoehn-Saric, "Normal physiological emotions but differences in expression of conscious feelings in children with high-functioning autism," *Journal of Autism and Developmental Disorders*, vol. 36, pp. 395–400, 2006.

[103] S. Berthoz and E. L. Hill, "The validity of using self-reports to assess emotion regulation abilities in adults with autism spectrum disorder," *European Psychiatry*, vol. 20, pp. 291–298, 2005.

[104] A. Sano, J. Hernandez, J. Deprey, M. Eckhardt, M. S. Goodwin, and R. W. Picard, "Multi- modal annotation tool for challenging behaviors in people with autism spectrum disor- ders," in *International Electronic Conference on Sensors and Applications*, ACM Conference on Ubiquitous Computing, 2012, pp. 737–740.

[105] A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders (DSM-5Æ)*. American Psychiatric Association, 2013.

[106] E. B. Foa, T. M. Keane, M. J. Friedman, and J. A. Cohen, *Effective Treatments for PTSD: Practice Guidelines from the International Society for Traumatic Stress Studies*. Guilford Press, 2008.

[107] R. Bradley, J. Greene, E. Russ, L. Dutra, and D. Westen, "A multidimensional meta-analysis of psychotherapy for ptsd," *American Journal of Psychiatry*, vol. 162, no. 2, pp. 214–227, 2005.

[108] A. Rizzo, J. Difede, B. O. Rothbaum, G. Reger, J. Spitalnick, J. Cukor, and R. Mclay, "Development and early evaluation of the virtual iraq/afghanistan exposure therapy system for combat-related ptsd," *Annals of the New York Academy of Sciences*, vol. 1208, no. 1.

[109] C. HolmgÂrd, G. N. Yannakakis, K. I. Karstoft, and H. S. Andersen, "Stress detection for ptsd via the startlemart game," in *Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 523–528.

[110] J. N. Hughes and O. Kwok, "Classroom engagement mediates the effect of teacher-Äìstudent support on elementary students' peer acceptance: A prospective analysis," *Journal of School Psychology*, vol. 43, no. 6, pp. 465 – 480, 2006.

[111] S. S. Darnell, "Engageme: A tool to simplify the conveyance of complicated data," in *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, 2014, pp. 359–362.

[112] J. A. Healey and R. W. Picard, "Detecting stress during real-world driving tasks using physiological sensors," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 2, pp. 156–166, 2005.

[113] C. Nass, I. Jonsson, H. Harris, B. Reaves, J. Endo, S. Brave, and L. Takayama, "Improving automotive safety by pairing driver emotion and car voice emotion," in *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, 2005, pp. 1973–1976.

[114] M.-K. Shan, F.-F. Kuo, M.-F. Chiang, and S.-Y. Lee, "Emotion-based music recommendation by affinity discovery from film music," *Expert systems with applications*, vol. 36, no. 4, pp. 7666–7674, 2009.

[115] M. Tkalčič, U. Burnik, and A. Košir, "Using affective parameters in a content-based recommender system for images," *User Modeling and User-Adapted Interaction*, vol. 20, no. 4, pp. 279–311, 2010.

[116] J. J. Kierkels, M. Soleymani, and T. Pun, "Queries and tags in affect-based multimedia retrieval," in *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*. IEEE, 2009, pp. 1436–1439.

[117] A. G. Money and H. Agius, *Feasibility of Personalized Affective Video Summaries*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 194–208.

[118] H. Katti, K. Yadati, M. Kankanhalli, and C. Tat-Seng, "Affective video summarization and story board generation using pupillary dilation and eye gaze," in *Multimedia (ISM), 2011 IEEE International Symposium on*, Dec 2011, pp. 319–326.

[119] E. Oliveira, P. Martins, and T. Chambel, "Ifelt: Accessing movies through our emotions," in *Proceddings of the 9th International Interactive Conference on Interactive Television*, ser. EuroITV '11. New York, NY, USA: ACM, 2011, pp. 105–114.

[120] M. Tkalcic, A. Odic, A. Kosir, and J. Tasic, "Affective labeling in a content-based recommender system for images," *IEEE transactions on multimedia*, vol. 15, no. 2, pp. 391–400, 2013.

[121] G. Acampora and A. Vitiello, "Interoperable neuro-fuzzy services for emotion-aware ambient intelligence," *Neurocomputing*, vol. 122, pp. 3–12, 2013.

[122] M. Hoque, M. Courgeon, J. Martin, B. Mutlu, and R. W. Picard, "Mach: My automated conversation coach," in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, 2013, pp. 697–706.

[123] P. Ekman and W. V. Friesen, *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, 1978.

[124] A. Ryan, J. F. Cohn, S. Lucey, J. Saragih, P. Lucey, F. D. la Torre, and A. Rossi, "Automated facial expression recognition system," in *43rd Annual 2009 International Carnahan Conference on Security Technology*, 2009, pp. 172–177.

[125] F. Negini, R. L. Mandryk, and K. G. Stanley, "Using affective state to adapt characters, npcs, and the environment in a first-person shooter game," in *Games Media Entertainment (GEM), 2014 IEEE*. IEEE, 2014, pp. 1–8.

[126] N. Bianchi-Berthouze, W. W. Kim, and D. Patel, "Does body movement engage you more in digital game play? and why?" in *International conference on affective computing and intelligent interaction*. Springer, 2007, pp. 102–113.

[127] N. Bianchi-Berthouze, "Understanding the role of body movement in player engagement," *Human–Computer Interaction*, vol. 28, no. 1, pp. 40–75, 2013.

[128] G. Colombetti, "From affect programs to dynamical discrete emotions," *Philosophical Psychology*, vol. 22, no. 4, pp. 407–425, 2009.

[129] P. Ekman, "An argument for basic emotions," *Cognition & emotion*, vol. 6, no. 3-4, pp. 169–200, 1992.

[130] W. Parrott, *Emotions in Social Psychology: Essential Readings*, ser. Key readings in social psychology. Psychology Press, 2001. [Online]. Available: https://books.google.it/books?id=jV5QVgM6Me8C

[131] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161–1178, 1980.

[132] R. Plutchik, "The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice," *American Scientist*, vol. 89, no. 4, pp. 344–350, 2001.

[133] A. Haag, S. Goronzy, P. Schaich, and J. Williams, "Emotion recognition using bio-sensors: First steps towards an automatic system," in *Tutorial and research workshop on affective dialogue systems*. Springer, 2004, pp. 36–48.

[134] P. J. Lang, M. M. Bradley, and B. N. Cuthbert, "International affective picture system (IAPS): Affective ratings of pictures and instruction manual," The Center for Research in Psychophysiology, University of Florida, Gainesville, FL, Tech. Rep. A-8, 2008. [Online]. Available: http://csea.phhp.ufl.edu/Media.html

[135] G. Chanel, K. Ciftci, J. C. Mota, A. Savran, L. H. Viet, L. Akarun, A. Caplier, M. Rombaut, and B. Sankur, "Emotion Detection in the Loop from Brain Signals and Facial Images," *Workshop on Emotion-Based Agent Architectures, Third International Conference on Autonomous Agents*, 2006.

[136] J. J. Gross and R. W. Levenson, "Emotion elicitation using films," *Cognition & emotion*, vol. 9, no. 1, pp. 87–108, 1995.

[137] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[138] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[139] J. Kim and E. André, "Emotion recognition based on physiological changes in music listening," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 12, pp. 2067–2083, 2008.

[140] M. Bradley, P. Lang, U. of Florida. Center for the Study of Emotion, Attention, and N. I. of Mental Health (Estats Units d'Amèrica), *The International affective digitized sounds (IADS)[: stimuli, instruction manual and affective ratings*. NIMH Center for the Study of Emotion and Attention, 1999. [Online]. Available: https://books.google.it/books?id=NFeZMwEACAAJ

[141] P. Rani, N. Sarkar, C. A. Smith, and J. A. Adams, "Affective communication for implicit human-machine interaction," in *Systems, Man and Cybernetics, 2003. IEEE International Conference on*, vol. 5, Oct 2003, pp. 4896–4903 vol.5.

[142] M. Mauri, V. Magagnin, P. Cipresso, L. Mainardi, E. N. Brown, S. Cerutti, M. Villamira, and R. Barbieri, "Psychophysiological signals associated with affective states," in *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology*, Aug 2010, pp. 3563–3566.

[143] F. Nasoz, K. Alvarez, C. L. Lisetti, and N. Finkelstein, "Emotion recognition from physiological signals using wireless sensors for presence technologies," *Cognition, Technology & Work*, vol. 6, no. 1, pp. 4–14, Feb 2004.

[144] C. Peter and A. Herbon, "Emotion representation and physiology assignments in digital systems," *Interacting with computers*, vol. 18, no. 2, pp. 139–170, 2006.

[145] J. Kim, "Bimodal emotion recognition using speech and physiological changes," in *Robust speech recognition and understanding*. InTech, 2007.

[146] H. Gunes and M. Piccardi, "A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior," in *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, vol. 1. IEEE, pp. 1148–1153.

[147] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[148] M. D. Robinson and G. L. Clore, "Episodic and semantic knowledge in emotional self-report: evidence for two judgment processes." *Journal of personality and social psychology*, vol. 83, no. 1, p. 198, 2002.

[149] I. B. Mauss and M. D. Robinson, "Measures of emotion: A review," *Cognition and emotion*, vol. 23, no. 2, pp. 209–237, 2009.

[150] A. Mehrabian and J. Russell, *An approach to environmental psychology*. M.I.T. Press, 1974.

[151] M. M. Bradley and P. J. Lang, "Measuring emotion: the self-assessment manikin and the semantic differential," *Journal of behavior therapy and experimental psychiatry*, vol. 25, no. 1, pp. 49–59, 1994.

[152] M. Kandemir, A. Vetek, M. Gönen, A. Klami, and S. Kaski, "Multi-task and multi-view learning of user state," *Neurocomputing*, vol. 139, pp. 97–106, 2014.

[153] J. A. Healey, "Wearable and automotive systems for affect recognition from physiology," Ph.D. dissertation, Massachusetts Institute of Technology, 2000.

[154] A. Kleinsmith and N. Bianchi-Berthouze, "Affective body expression perception and recognition: A survey," *IEEE Transactions on Affective Computing*, vol. 4, no. 1, pp. 15–33, 2013.

[155] C. Busso and S. S. Narayanan, "Interrelation between speech and facial gestures in emotional utterances: A single subject study," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2331–2347, 2007.

[156] A. Chakraborty, A. Konar, U. K. Chakraborty, and A. Chatterjee, "Emotion recognition from facial expressions and its control using fuzzy logic," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 39, no. 4, pp. 726–743, 2009.

[157] R. Cowie, E. Douglas-Cowie, J. G. Taylor, S. Ioannou, M. Wallace, and S. Kollias, "An intelligent system for facial emotion recognition," in *2005 IEEE International Conference on Multimedia and Expo*, 2005.

[158] M. Valstar and M. Pantic, "Fully automatic facial action unit detection and temporal analysis," pp. 149–149, 2006.

[159] Z. Liu, M. Wu, W. Cao, L. Chen, J. Xu, R. Zhang, M. Zhou, and J. Mao, "A facial expression emotion recognition based human-robot interaction system," *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 4, pp. 668–676, 2017.

[160] R. W. Picard, E. Vyzas, and J. Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.

[161] Y. Zhang, Z. Li, F. Ren, and S. Kuroiwa, "Semi-automatic emotion recognition from textual input based on the constructed emotion thesaurus," in *2005 International Conference on Natural Language Processing and Knowledge Engineering*, 2005, pp. 571–576.

[162] H. Li and F. Ren, "The study on text emotional orientation based on a three-dimensional emotion space model," in *2009 International Conference on Natural Language Processing and Knowledge Engineering*, 2009, pp. 1–6.

[163] K. Matsumoto, J. Minato, F. Ren, and S. Kuroiwa, "Estimating human emotions using wording and sentence patterns," in *2005 IEEE International Conference on Information Acquisition*, 2005.

[164] R. A. Calix, S. A. Mallepudi, B. Chen, and G. M. Knapp, "Emotion recognition in text for 3-d facial expression rendering," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 544–551, 2010.

[165] G. Riva, F. Davide, and W. A. Ijsselsteijn, *Being There: Concepts, effects and measurement of user presence in [35] synthetic environments*. IOS Press, 2003.

[166] T. Danisman and A. Alpkocak, "Feeler: emotion classification of text using vector space model," in *Proceedings of the AISB 2008 Symposium on Affective*, vol. 2, 2008, pp. 53–59.

[167] V. K. Jain, S. Kumar, and S. L. Fernandes, "Extraction of emotions from multilingual text using intelligent text processing and computational linguistics," *Journal of Computational Science*, vol. 21, pp. 316–326, 2017.

[168] J. H. Jeon, R. Xia, and Y. Liu, "Sentence level emotion recognition based on decisions from subsentence segments," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4940–4943.

[169] S. G. Koolagudi and K. S. Rao, "Real life emotion classification using vop and pitch based spectral features," in *2010 Annual IEEE India Conference (INDICON)*, 2010, pp. 1–4.

[170] A. A. Razak, R. Komiya, M. Izani, and Z. Abidin, "Comparison between fuzzy and nn method for speech emotion recognition," in *Third International Conference on Information Technology and Applications (ICITA'05)*, vol. 1, 2005, pp. 297–302.

[171] T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "A dimensional approach to emotion recognition of speech from movies," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2009, pp. 65–68.

[172] S. G. Koolagudi, R. Reddy, and K. S. Rao, "Emotion recognition from speech signal using epoch parameters," in *International Conference on Signal Processing and Communications (SPCOM)*, 2010, pp. 1–5.

[173] S. Tokuno, G. Tsumatori, S. Shono, E. Takei, T. Yamamoto, G. Suzuki, S. Mituyoshi, and M. Shimura, "Usage of emotion recognition in military health care," in *2011 Defense Science Research Conference and Expo (DSR)*, 2011, pp. 1–5.

[174] H. G. Wallbott, "Bodily expression of emotion," *European Journal of Social Psychology*, vol. 28, no. 6, 1998.

[175] G. Castellano, S. D. Villalba, and A. Camurri, "Recognising human emotions from body movement and gesture dynamics," in *Affective Computing and Intelligent Interaction*. Springer Berlin Heidelberg, 2007, pp. 71–82.

[176] A. Kapur, A. Kapur, N. Virji-Babul, G. Tzanetakis, and P. F. Driessen, "Gesture-based affective computing on motion capture data," in *Affective Computing and Intelligent Interaction*. Springer Berlin Heidelberg, 2005, pp. 1–7.

[177] A. Kleinsmith and N. Bianchi-Berthouze, "Recognizing affective dimensions from body posture," *Affective computing and intelligent interaction*, pp. 48–58, 2007.

[178] J. Wagner, J. Kim, and E. André, "From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE, 2005, pp. 940–943.

[179] E. van den Broek, J. Janssen, J. Healey, and M. van der Zwaag, *Prerequisites for Affective Signal Processing (ASP) - Part II*. INSTICC PRESS, 1 2010, pp. 188–193.

[180] C. Peter, E. Ebert, and H. Beikirch, "Physiological sensing for affective computing," in *Affective Information Processing*. Springer, 2009, pp. 293–310.

[181] E. L. Broek, *Affective Signal Processing (ASP): Unraveling the mystery of emotions*. University of Twente, 2011.

[182] M. Brennan, M. Palaniswami, and P. Kamen, "Do existing measures of poincare plot geometry reflect nonlinear features of heart rate variability?" *IEEE transactions on biomedical engineering*, vol. 48, no. 11, pp. 1342–1347, 2001.

[183] D. Giakoumis, D. Tzovaras, K. Moustakas, and G. Hassapis, "Automatic recognition of boredom in video games using novel biosignal moment-based features," *IEEE Transactions on Affective Computing*, vol. 2, no. 3, pp. 119–133, July 2011.

[184] F. Liu, G. Liu, and X. Lai, "Emotional intensity evaluation method based on galvanic skin response signal," in *Computational Intelligence and Design (ISCID), 2014 Seventh International Symposium on*, vol. 1. IEEE, 2014, pp. 257–261.

[185] E. van den Broek, J. Janssen, J. Westerink, and J. Healey, *Prerequisites for affective signal processing (ASP)*. INSTICC PRESS, 1 2009, pp. 426–433.

[186] S. Greene, H. Thapliyal, and A. Caban-Holt, "A survey of affective computing for stress detection: Evaluating technologies in stress detection for better health," *IEEE Consumer Electronics Magazine*, vol. 5, no. 4, pp. 44–56, 2016.

[187] A. Alberdi, A. Aztiria, and A. Basarab, "Towards an automatic early stress recognition system for office environments based on multimodal measurements: A review," *Journal of biomedical informatics*, vol. 59, pp. 49–75, 2016.

[188] N. Sharma and T. Gedeon, "Objective measures, sensors and computational techniques for stress recognition and classification: A survey," *Computer methods and programs in biomedicine*, vol. 108, no. 3, pp. 1287–1301, 2012.

[189] K. H. Kim, S. W. Bang, and S. R. Kim, "Emotion recognition system using short-term monitoring of physiological signals," *Medical and biological engineering and computing*, vol. 42, no. 3, pp. 419–427, 2004.

[190] C. L. Lisetti and F. Nasoz, "Using noninvasive wearable computers to recognize human emotions from physiological signals," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 11, p. 929414, 2004.

[191] O. Villon and C. Lisetti, "A user model of psycho-physiological measure of emotion," in *International Conference on User Modeling*. Springer, 2007, pp. 319–323.

[192] G. N. Yannakakis and J. Hallam, "Entertainment modeling through physiology in physical play," *International Journal of Human-Computer Studies*, vol. 66, no. 10, pp. 741–755, 2008.

[193] L. Shen, V. Callaghan, and R. Shen, "Affective e-learning in residential and pervasive computing environments," *Information Systems Frontiers*, vol. 10, no. 4, pp. 461–472, 2008.

[194] C.-Y. Chang, J.-S. Tsai, C.-J. Wang, and P.-C. Chung, "Emotion recognition with consideration of facial expression and physiological signals," in *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*. IEEE, 2009, pp. 278–283.

[195] F. Nasoz, C. L. Lisetti, and A. V. Vasilakos, "Affectively intelligent and adaptive car interfaces," *Information Sciences*, vol. 180, no. 20, pp. 3817–3836, 2010.

[196] A. de Santos Sierra, C. S. Ávila, G. B. del Pozo, and J. G. Casanova, "Stress detection by means of stress physiological template," in *2011 Third World Congress on Nature and Biologically Inspired Computing*, Oct 2011, pp. 131–136.

[197] A. de Santos Sierra, C. S. Avila, J. G. Casanova, and G. B. del Pozo, "A stress-detection system based on physiological signals and fuzzy logic," *IEEE Transactions on Industrial Electronics*, vol. 58, no. 10, pp. 4857–4865, Oct 2011.

[198] P. Melillo, M. Bracale, and L. Pecchia, "Nonlinear heart rate variability features for real-life stress detection. case study: students under stress due to university examination," *Biomedical engineering online*, vol. 10, no. 1, p. 96, 2011.

[199] J. Hernandez, R. R. Morris, and R. W. Picard, "Call center stress recognition with person-specific models," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2011, pp. 125–134.

[200] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, "Ecg pattern analysis for emotion detection," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 102–115, 2012.

[201] K. Palanisamy, M. Murugappan, and S. Yaacob, "Multiple physiological signal-based human stress identification using non-linear classifiers," *Elektronika ir elektrotechnika*, vol. 19, no. 7, pp. 80–85, 2013.

[202] A. Sano and R. W. Picard, "Stress recognition using wearable sensors and mobile phones," in *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 2013, pp. 671–676.

[203] X. Li, Z. Chen, Q. Liang, and Y. Yang, "Analysis of mental stress recognition and rating based on hidden markov model," in *Journal of Computational Information Systems*, 2014, pp. 7911–7919.

[204] W. Wen, G. Liu, N. Cheng, J. Wei, P. Shangguan, and W. Huang, "Emotion recognition based on multi-variant correlation of physiological signals," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 126–140, April 2014.

[205] G. Valenza, M. Nardelli, A. Lanata, C. Gentili, G. Bertschy, R. Paradiso, and E. P. Scilingo, "Wearable monitoring for mood recognition in bipolar disorder based on history-dependent long-term heart rate variability analysis," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 5, pp. 1625–1635, 2014.

[206] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo, "Recognizing emotions induced by affective sounds through heart rate variability," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 385–394, 2015.

[207] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, "Continuous stress detection using a wrist device: in laboratory and real life," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, 2016, pp. 1185–1193.

[208] W. S. Liew, M. Seera, C. K. Loo, E. Lim, and N. Kubota, "Classifying stress from heart rate variability using salivary biomarkers as reference," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 10, pp. 2035–2046, 2016.

[209] L. Ciabattoni, F. Ferracuti, S. Longhi, L. Pepa, L. Romeo, and F. Verdini, "Real-time mental stress detection based on smartwatch," in *Consumer Electronics (ICCE), 2017 IEEE International Conference on*. IEEE, 2017, pp. 110–111.

[210] M. Wöllmer, F. Eyben, S. Reiter, B. Schuller, C. Cox, E. Douglas-Cowie, and R. Cowie, "Abandoning emotion classes-towards continuous emotion recognition with modelling of long-range dependencies," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[211] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[212] ——, "Output-associative rvm regression for dimensional and continuous emotion prediction," *Image and Vision Computing*, vol. 30, no. 3, pp. 186 – 196, 2012, best of Automatic Face and Gesture Recognition 2011.

[213] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artificial intelligence*, vol. 89, no. 1, pp. 31–71, 1997.

[214] T. Tong, R. Wolz, Q. Gao, R. Guerrero, J. V. Hajnal, D. Rueckert, A. D. N. Initiative *et al.*, "Multiple instance learning for classification of dementia in brain mri," *Medical image analysis*, vol. 18, no. 5, pp. 808–818, 2014.

[215] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Advances in neural information processing systems*, 2003, pp. 577–584.

[216] A. Zafra, C. Romero, and S. Ventura, "Multiple instance learning for classifying students in learning management systems," *Expert Systems with Applications*, vol. 38, no. 12, pp. 15 020–15 031, 2011.

[217] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 8, pp. 1619–1632, 2011.

[218] T. Rao, M. Xu, H. Liu, J. Wang, and I. Burnett, "Multi-scale blocks based image emotion classification using multiple instance learning," in *Image Processing (ICIP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 634–638.

[219] K. Sikka, A. Dhall, and M. Bartlett, "Weakly supervised pain localization using multiple instance learning," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[220] B. Wu, E. Zhong, A. Horner, and Q. Yang, "Music emotion recognition by multi-label multi-layer multi-instance multi-view learning," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 117–126.

[221] J. Gibson, A. Katsamanis, F. Romero, B. Xiao, P. Georgiou, and S. Narayanan, "Multiple instance learning for behavioral coding," *IEEE Transactions on Affective Computing*, 2015.

[222] O. Maron and T. Lozano-Pérez, "A framework for multiple-instance learning," in *Advances in neural information processing systems*, 1998, pp. 570–576.

[223] Q. Zhang and S. A. Goldman, "Em-dd: An improved multiple-instance learning technique," in *Advances in neural information processing systems*, 2002, pp. 1073–1080.

[224] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola, "Multi-instance kernels," in *In Proc. 19th International Conf. on Machine Learning*. Morgan Kaufmann, 2002, pp. 179–186.

[225] R. C. Bunescu and R. J. Mooney, "Multiple instance learning for sparse positive bags," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 105–112.

[226] P. J. Lang, M. K. Greenwald, M. M. Bradley, and A. O. Hamm, "Looking at pictures: Affective, facial, visceral, and behavioral reactions," *Psychophysiology*, vol. 30, no. 3, pp. 261–273, 1993.

[227] C. L. Lisetti and F. Nasoz, "Using noninvasive wearable computers to recognize human emotions from physiological signals," *EURASIP Journal on Advances in Signal Processing*, vol. 2004, no. 11, p. 929414, 2004.

[228] G. Chanel, J. J. Kierkels, M. Soleymani, and T. Pun, "Short-term emotion assessment in a recall paradigm," *International Journal of Human-Computer Studies*, vol. 67, no. 8, pp. 607–627, 2009.

[229] W. Wen, G. Liu, N. Cheng, J. Wei, P. Shangguan, and W. Huang, "Emotion recognition based on multi-variant correlation of physiological signals," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 126–140, 2014.

[230] M. Nardelli, G. Valenza, A. Greco, A. Lanata, and E. P. Scilingo, "Recognizing emotions induced by affective sounds through heart rate variability," *IEEE Transactions on Affective Computing*, vol. 6, no. 4, pp. 385–394, 2015.

[231] E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. Mcrorie, J.-C. Martin, L. Devillers, S. Abrilian, A. Batliner *et al.*, "The humaine database: addressing the collection and annotation of naturalistic and induced emotional data," *Affective computing and intelligent interaction*, pp. 488–500, 2007.

[232] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[233] G. N. Yannakakis and J. Hallam, "Entertainment modeling in physical play through physiology beyond heart-rate," in *International Conference on Affective Computing and Intelligent Interaction*. Springer, 2007, pp. 254–265.

[234] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, "Ecg pattern analysis for emotion detection," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 102–115, Jan 2012.

[235] J. Wang and Y. Gong, "Recognition of multiple drivers' emotional state," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[236] L. Ciabattoni, F. Ferracuti, S. Longhi, L. Pepa, L. Romeo, and F. Verdini, "Multimedia experience enhancement through affective computing," in *Consumer Electronics (ICCE), 2017 IEEE International Conference on*. IEEE, 2017, pp. 182–183.

[237] R. Cai, C. Zhang, C. Wang, L. Zhang, and W.-Y. Ma, "Musicsense: contextual music recommendation using emotional allocation modeling," in *Proceedings of the 15th ACM international conference on Multimedia*. ACM, 2007, pp. 553–556.

[238] L. Lu, D. Liu, and H.-J. Zhang, "Automatic mood detection and tracking of music audio signals," *IEEE Transactions on audio, speech, and language processing*, vol. 14, no. 1, pp. 5–18, 2006.

[239] E. M. Schmidt and Y. E. Kim, "Prediction of time-varying musical mood distributions using kalman filtering," in *Machine Learning and Applications (ICMLA), 2010 Ninth International Conference on*. IEEE, 2010, pp. 655–660.

[240] H. Gunes and M. Pantic, "Automatic, dimensional and continuous emotion recognition," *Int. J. Synth. Emot.*, vol. 1, no. 1, pp. 68–99, Jan. 2010.

[241] J. D. Velásquez, "Modeling emotions and other motivations in synthetic agents," in *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Conference on Innovative Applications of Artificial Intelligence*, ser. AAAI'97/IAAI'97, 1997, pp. 10–15.

[242] D. Hailey, R. Roine, A. Ohinmaa, and L. Dennett, "Evidence of benefit from telerehabilitation in routine care: a systematic review," *Journal of Telemedicine and Telecare*, vol. 17, no. 6, pp. 281–287, 2011.

[243] K. Steel, D. Cox, and H. Garry, "Therapeutic videoconferencing interventions for the treatment of long-term conditions," *Journal of Telemedicine and Telecare*, vol. 17, no. 3, pp. 109–117, 2011.

[244] M. S. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, M. Valstar, H. Meng, A. Kemp, M. Shafizadeh *et al.*, "The automatic detection of chronic pain-related expression: requirements, challenges and the multimodal emopain dataset," *IEEE transactions on affective computing*, vol. 7, no. 4, pp. 435–451, 2016.

[245] D. C. Turk *et al.*, "Iasp taxonomy of chronic pain syndromes: preliminary assessment of reliability," *Pain*, vol. 30, no. 2, pp. 177–189, 1987.

[246] J. W. Vlaeyen and S. J. Linton, "Fear-avoidance and its consequences in chronic musculoskeletal pain: a state of the art," *Pain*, vol. 85, no. 3, pp. 317–332, 2000.

[247] F. J. Keefe and A. R. Block, "Development of an observation method for assessing pain behavior in chronic low back pain patients," *Behavior Therapy*, vol. 13, no. 4, pp. 363–375, 1982.

[248] T. A. Olugbade, M. Aung, N. Bianchi-Berthouze, N. Marquardt, and A. C. Williams, "Bi-modal detection of painful reaching for chronic pain rehabilitation systems," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 455–458.

[249] H. Gonzalez-Jorge, P. Rodríguez-Gonzálvez, J. Martínez-Sánchez, D. González-Aguilera, P. Arias, M. Gesto, and L. Díaz-Vilariño, "Metrological comparison between kinect i and kinect ii sensors," *Measurement*, vol. 70, pp. 21–26, 2015.

[250] X. Xu, R. W. McGorry, L.-S. Chou, J. hua Lin, and C. chi Chang, "Accuracy of the microsoft kinect^{TM} for measuring gait parameters during treadmill walking," *Gait and Posture*, vol. 42, no. 2, 2015.

[251] E. Dolatabadi, B. Taati, and A. Mihailidis, "Concurrent validity of the microsoft kinect for windows v2 for measuring spatiotemporal gait parameters," *Medical Engineering & Physics*, vol. 38, no. 9, pp. 952 – 958, 2016.

[252] B. F. Mentiplay, L. G. Perraton, K. J. Bower, Y.-H. Pua, R. McGaw, S. Heywood, and R. A. Clark, "Gait assessment using the microsoft xbox one kinect: Concurrent validity and inter-day reliability of spatiotemporal and kinematic variables," *Journal of biomechanics*, vol. 48, no. 10, pp. 2166–2170, 2015.

[253] X. Xu and R. W. McGorry, "The validity of the first and second generation Microsoft Kinect^{TM} for identifying joint center locations during static postures," *Applied Ergonomics*, vol. 49, Jul. 2015.

[254] B. Galna, G. Barry, D. Jackson, D. Mhiripiri, P. Olivier, and L. Rochester, "Accuracy of the microsoft kinect sensor for measuring movement in people with parkinson's disease," *Gait and Posture*, vol. 39, no. 4, 2014.

[255] A. Schmitz, M. Ye, R. Shapiro, R. Yang, and B. Noehren, "Accuracy and repeatability of joint angles measured using a single camera markerless motion capture system," *Journal of biomechanics*, vol. 47, no. 2, pp. 587–591, 2014.

[256] L. R. Reither, M. H. Foreman, N. Migotsky, C. Haddix, and J. R. Engsberg, "Upper extremity movement reliability and validity of the kinect version 2," *Disability and Rehabilitation: Assistive Technology*, vol. 0, no. 0, pp. 1–9, 0.

[257] A. Mobini, S. Behzadipour, and M. S. Foumani, "Accuracy of kinect's skeleton tracking for upper body rehabilitation applications," *Disability and Rehabilitation: Assistive Technology*, vol. 9, no. 4, pp. 344–352, 2014.

[258] A. de Albuquerque, E. Moura, T. Vasconcelos, L. Mendes, and D. Nagem, "Kinect sensor used as a support tool in clinical analysis," *Journal of Biomechanics*, vol. 45, p. S304, 2012.

[259] T. W. Macpherson, J. Taylor, T. McBain, M. Weston, and I. R. Spears, "Real-time measurement of pelvis and trunk kinematics during treadmill locomotion using a low-cost depth-sensing camera: A concurrent validity study," *Journal of biomechanics*, vol. 49, no. 3, pp. 474–478, 2016.

[260] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, S. Longhi, L. Romeo, and F. V. S. N. Russi, "Accuracy evaluation of the kinect v2 sensor during dynamic movements in a rehabilitation scenario," in *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug 2016, pp. 8034–8037.

[261] A. Mishra, M. Skubic, and C. Abbott, "Development and preliminary validation of an interactive remote physical therapy system," in *EMBC, 37th Annual Int. Conf. of the IEEE*, Aug 2015.

[262] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, V. Kyrki, S. Longhi, L. Romeo, and F. Verdini, "Physical rehabilitation exercises assessment based on hidden semi-markov model by kinect v2," in *IEEE-EMBS Int. Conf. on BHI*, Feb 2016, pp. 256–259.

[263] G. A. of the World Medical Association *et al.*, "World medical association declaration of helsinki: ethical principles for medical research involving human subjects." *The Journal of the American College of Dentists*, vol. 81, no. 3, p. pp. 14, 2014.

[264] C. Kisner and L. Colby, *Therapeutic Exercise: Foundations and Techniques*. F.A. Davis, 2012, last accessed: 2017/02/23. [Online]. Available: https://books.google.it/books?id=udY-AAAAQBAJ

[265] M. Hutson and A. Ward, *Oxford Textbook of Musculoskeletal Medicine*, Oxford, UK, 2015. [Online]. Available: http://oxfordmedicine.com/view/10.1093/med/9780199674107.001.0001/med-9780199674107

[266] "Microsoft band sdk." [Online]. Available: https://developer.microsoftband.com

[267] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, V. Kyrki, A. Monteriù, L. Romeo, and F. Verdini, "A hidden semi-markov model based approach for rehabilitation exercise assessment," *Journal of Biomedical Informatics*, vol. 78, pp. 1 – 11, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1532046417302824

[268] M. Capecci, M. G. Ceravolo, F. Ferracuti, S. Iarlori, V. Kyrki, S. Longhi, L. Romeo, and F. Verdini, "Physical rehabilitation exercises assessment based on hidden semi-markov model by kinect v2," in *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, Feb 2016, pp. 256–259.

[269] M. Pomplun, "Evaluation metrics and results of human arm movement imitation," in *In Proc. of First IEEE-RAS International Conference on Humanoid Robots*. Citeseer, 2000.

[270] S.-Z. Yu, "Hidden semi-markov models," *Artificial Intelligence*, vol. 174, no. 2, pp. 215 – 243, 2010.

[271] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[272] S.-Z. Yu and H. Kobayashi, "Practical implementation of an efficient forward-backward algorithm for an explicit-duration hidden markov model," *IEEE Trans. Signal Process.*, vol. 54, pp. 1947–1951, 2006.

[273] S. Calinon, A. Pistillo, and D. Caldwell, "Encoding the time and space constraints of a task in explicit-duration hidden markov model," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).*, 2011, pp. 3413–3418.

[274] M. Soleymani, F. Villaro-Dixon, T. Pun, and G. Chanel, "Toolbox for emotional feature extraction from physiological signals (teap)," *Frontiers in ICT*, vol. 4, 2017.

[275] O. Maron, "Learning from ambiguity," *Ph.D. dissertation, Massachusetts Inst. Techno., Cambridge, MA, USA*, 1998.

[276] M. Müller, *Information Retrieval for Music and Motion*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007.

[277] X. Long, P. Fonseca, J. Foussier, R. Haakma, and R. M. Aarts, "Sleep and wake classification with actigraphy and respiratory effort using dynamic warping," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. pp. 1272–1284, July 2014.

[278] A. Vakanski, H.-p. Jun, D. Paul, and R. Baker, "A data set of human body movements for physical rehabilitation exercises," *Data*, vol. 3, no. 1, 2018.

[279] D. Leightley, M. H. Yap, J. Coulson, Y. Barnouin, and J. S. McPhee, "Benchmarking human motion analysis using kinect one: An open source dataset," *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, pp. 1–7, 2015.

[280] A. Hanjalic and L.-Q. Xu, "Affective video content representation and modeling," *IEEE Transactions on multimedia*, vol. 7, no. 1, pp. 143–154, 2005.

[281] P. Lucey, J. F. Cohn, K. M. Prkachin, P. E. Solomon, and I. Matthews, "Painful data: The unbc-mcmaster shoulder pain expression archive database," in *Automatic Face & Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on*. IEEE, 2011, pp. 57–64.

[282] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li, "Multi-instance multi-label learning," *Artificial Intelligence*, vol. 176, no. 1, pp. 2291–2320, 2012.

[283] S. Ray and D. Page, "Multiple instance regression," in *Proceedings of the Eighteenth International Conference on Machine Learning*, ser. ICML '01, 2001, pp. 425–432.

[284] D. R. Dooly, Q. Zhang, S. A. Goldman, and R. A. Amar, "Multiple-instance learning of real-valued data," *Journal of Machine Learning Research*, vol. 3, no. Dec, pp. 651–678, 2002.

[285] F. Herrera, S. Ventura, R. Bello, C. Cornelis, A. Zafra, D. Sánchez-Tarragó, and S. Vluymans, "Multi-instance regression," in *Multiple Instance Learning*. Springer, 2016, pp. 127–140.