SCUOLA DI DOTTORATO
UNIVERSITÀ DEGLI STUDI DI MILANO-BICOCCA

Dipartimento di / Department of

Informatics, Systems and Communication

Dottorato di Ricerca in / PhD program: Computer Science   Ciclo / Cycle: XXXIII

Deep Psychology Recognition Based on Automatic Analysis of Non-Verbal Behaviors

Cognome / Surname: Khalifa   Nome / Name: Intissar

Matricola / Registration number: 836548

Supervisor: Prof. Mourad Zaied

Supervisor: Prof. Raimondo Schettini

Co-Supervisor: Dr. Ridha Ejbali

Coordinatore / Coordinator: Prof. Leonardo Mariani

ANNO ACCADEMICO / ACADEMIC YEAR 2019-2020

Dedications

I dedicate this milestone in my life to all those precious people.
***To my dear parents Mohamed and Moufida** who have been the symbol of love, tenderness, sacrifice, happiness, peace, and source of encouragement and inspiration to me throughout my life. Thank you for your endless patience, moral support and your permanent valuable pieces of advice that have always guided my steps towards success. I will always do my best to let you feel proud and never disappoint you. May God preserve you and grant you health and happiness.*
***To my sisters Achwak and Amira**, this work is a sign of my attachment and love, we shared the most pleasant and difficult moments. You have always been by my side to help me reach my objectives.*
***To my lovely aunt Noura** who has been a source of motivation, love, and strength during moments of despair and discouragement. May God protect you from harm.*
***To my big family**, I dedicate you this work as an expression of gratitude for your encouragement and support during the difficult moments of my life.*
To all those dear people I can never forget, I dedicate this work.

INTISSAR...

Acknowledgements

First of all, I want to thank my supervisor Prof. Mourad Zaied. I am very grateful to him for his tolerance in letting me to pursue my own academic path since without his counsel and early interest I would not have done any research at all. He provides me a lot of academic opportunities that would not have otherwise been possible. His insight and feedback have been invaluable.

I want also to thank my supervisor Prof. Raimondo Schettini. I am honored to work with him. Despite all his occupations, he always found time to listen, discuss and provide guidance. During my period abroad, he has always been by my side to help and support me in very difficult moments.

I would like to express my gratitude to my co-supervisor Prof. Ridha Ejbali for his considerable time and availability of mind. He has been helpful in providing advice, precious suggestions, and insightful guidance.

I would like to express my deep recognition to committee members for the acceptation to evaluate my thesis and for the considerable time, energy, and availability of mind.

I also want to extend my thanks to all my colleagues and the doctors at the RTIM research unit, the beautiful times we spent together especially in meetings and workshops will remain firmly in my mind.

I want to thank my friends in the Imaging and Vision Laboratory, I had the chance to discuss and reason about anything. I shared many pleasant moments, I never felt lonely despite my distance from my country and my family.

Special thanks to any person in the Department of Informatics, Systems and Communication who helped me in very critical moments.

"When dealing with people, remember you are not dealing with creatures of logic, but with creatures of emotion."
Dale Carnegie

"The only way to change someone's mind is to connect with them from the heart."
Rasheed Ogunlaru

Sommario

Un aspetto estremamente cruciale nel dominio dell'interazione uomo-uomo è la comunicazione delle emozioni. Essere in grado di dedurre gli stati emotivi attraverso comportamenti non-verbali consente agli esseri umani di comprendere e ragionare su obiettivi ed intenti altrui. L'Affective Computing è una branca dell'informatica che mira a trarre vantaggio dal potere delle emozioni per facilitare un'interazione uomo-macchina più efficiente. L'obiettivo è dare alle macchine la capacità di esprimere, riconoscere e regolare le emozioni. In questa tesi, esamineremo in dettaglio il ruolo delle espressioni visive ed uditive nel comunicare emozioni, e svilupperemo modelli computazionali per il riconoscimento automatico delle emozioni: un'area di ricerca molto attiva nell'ultimo decennio. In generale, la comunicazione delle emozioni attraverso i segnali del corpo è compresa in misura minore rispetto a altre modalità. La psicologia sociale che ha ispirato molti approcci computazionali si è tradizionalmente concentrata sui segnali facciali. Tuttavia, la gestualità del corpo è una fonte significativa di informazioni, soprattutto quando altri canali sono nascosti o in presenza di sottili sfumature di espressioni. In questo contesto, proporremo diversi approcci per il riconoscimento di gesti con applicazione alle emozioni, utilizzando due modelli. Per il modello basato su parti, svilupperemo un approccio ibrido che incorpora due tecniche di stima del movimento e di normalizzazione temporale per la modellazione del movimento della mano. Passeremo poi a presentare il nostro approccio spazio-temporale profondo (deep) per modellare il movimento del corpo, ed infine ottenere lo stato emotivo della persona. In questa parte, dimostrereamo che la nostra tecnica basata sul deep learning supera le tradizionali tecniche di machine learning. Per il modello basato sulla cinematica, combineremo la stima della posa del soggetto (con applicazione al rilevamento dello scheletro) e la classificazione delle emozioni per proporre una nuova architettura profonda a più stadi in grado di affrontare entrambi i compiti sfruttando i punti di forza dei modelli pre-addestrati. Dimostreremo che le tecniche di transfer learning superano le tradizionali tecniche di apprendimento automatico. Come ulteriore modalità, il parlato è la forma più comune e veloce per comunicare tra esseri umani. Questa realtà ci ha spinti a riconoscere le condizioni emotive del soggetto parlante in maniera automatica tramite la sua voce. Proporremo una rappresentazione profonda di tipo temporale e basata

sul cepstrum, che sfrutta la concatenazione di feature spettrali, feature di basate su derivate temporali, ed un classificatore basato sul deep learning per il riconoscimento delle emozioni del parlato. I risultati ottenuti per entrambe le modalità utilizzando i nostri metodi sono molto promettenti e competitivi rispetto ai metodi esistenti nello stato dell'arte. Riteniamo che il nostro lavoro sia pertinente sia per il social computing che per la psicologia organizzativa. Prendendo come esempio i colloqui di lavoro, un ambito ben studiato dagli psicologi sociali, il nostro studio può fornire informazioni utili su come sfruttare i segnali non verbali per supportare le aziende nel processo di assunzione. Questa tesi descrive la fattibilità di usare indizi estratti automaticamente per analizzare gli stati psicologici, come interessante alternativa alle annotazioni manuali dei segnali comportamentali.
Parole chiave: Comportamenti non-verbali, emozione, gesti del corpo, linguaggio parlato, modello basato su parti, modello basato su cinematica.

Abstract

One highly crucial aspect in the domain of human-human interaction is the communication of emotions. Being able to deduce emotional states through non-verbal behaviors allows humans to understand and reason about each others' underlying goals and intents. Affective Computing is the branch of computer science that aims to profit from the power of emotions to facilitate a more efficient human-machine interaction. The goal is to give the machines the ability to express, recognize, and regulate emotions. In this dissertation, we look in detail at the role of visual and auditory expressions for communicating emotions and we develop computational models for automatic emotion recognition which is an active research area over the last decade. In general, communication of emotions through body cues is less understood than other modalities. Social psychology that has inspired many computational approaches has traditionally focused on facial cues. However, body gestures are a significant source of information especially when other channels are hidden or there is a subtle nuance of expressions. In this context, we propose our approaches for emotional body gesture recognition using two different models. For the part-based model, we develop a hybrid approach that incorporates two techniques of motion estimation and temporal normalization for hand motion modeling, then we move to present our deep-spatio temporal approach for body motion modeling to have finally the person's emotional state. In this part, we demonstrate that our deep learning technique outperforms traditional machine learning techniques. For the kinematic-based model, we combine human pose estimation for skeleton detection and emotion classification to propose a new deep multi-stage architecture able to deal with both tasks by exploiting the strong points of models pre-trained. We demonstrate that transfer learning techniques outperform traditional machine learning techniques. As another modality, speech is the fastest normal way to communicate among human. This reality motivates us to identify the emotional conditions of the uttering person by utilizing his/her voice automatically. We propose a deep temporal-cepstrum representation based on the concatenation of spectral features, temporal derivatives features, and a deep learning classifier for speech emotion recognition. The results obtained for both modalities using our suggested methods are very promising and competitive over existing methods in the state of the art. We believe that our work is pertinent for both social computing and organizational psychology. Taking the example of job interviews, which is well studied by social psychologists, our study may provide insights for how non-verbal cues could be used by the companies for the hiring decision. In fact, our dissertation shows the feasibility of using automatically extracted cues to analyze the psychological states as an attractive alternative to manual annotations of behavioral cues.

Keywords: Non-verbal behaviors, emotion, body gestures, speech, part-based model, kinematic-based model.

Contents

List of Figures

List of Tables

Acronyms and abbreviations

We enumerate here abbreviations and acronyms recommended and mentioned in this report.

AC	Affective Computing
AI	Artificial Intelligence
AE	Auto-Encoder
BN	Bayes Network
BOW	Bag of Words
BEL	Brain Emotional Learning model
BP	Back-Propagation
COCO	Common Object in Context
CNN	Convolutional Neural Network
CCCNN	Cross-Channel Convolutional Neural Network
DT	Decision Tree
DTW	Dynamic Time Warping
DBN	Deep Belief Network
DWT	Discrete Wavelet Transform
DCN	Depthwise Convolutional Network
DSC	Depthwise Separable Convolution
DCT	Discrete Cosine Transform
ET	Ensemble Tree
EBMI	Energy Binary Motion Image
FFT	Fast Fourier Transform
GMM	Gaussian Mixture Model
GNB	Gaussian Naive Bayes
GPU	Graphics Processing Unit
HMM	Hidden Markov Model
HMI	History Motion Image
HOG	Histogram of Oriented Gradients
HNB	Hidden NaÃ¯ve Bayes
HSOF	Horn Schunck Optical Flow
HPI	History Pose Image

KF	Kalman Filter
KNN	K-Nearest Neighbor
LSTM	Long Short-Term Memory
LPC	Linear Predictive Coding
LMT	Logistic Model Tree
MCCNN	Multi-Channel Convolutional Neural Network
MFCC	Mel-Frequency Cepstral Coefficients
MDS	Multi-Dimensional Scaling
NN	Neural Network
OF	Optical Flow
RF	Random Forest
ReLU	Rectified Linear Unit
SER	Speech Emotion Recognition
SVM	Support Vector Machine
SVC	Support Vector Classifier
STE	Short Term Energy
SSAE	Stacked Sparse Auto-Encoder
SAE	Sparse Auto-Encoder
SC	Softmax Classifier
SGD	Stochastic Gradient Descent
TN	Temporal Normalization
TEO	Teager Energy Operator
VGG	Visual Geometry Group
ZCR	Zero Crossing Rate

Chapter 1

Introduction

1.1 Motivations and problem statement

Emotions color our lives, allow us to express different facets of our personality. In the previous century, there was a lot of belief in the omnipotence of reason, forgetting emotion. It was considered an obstacle to the work of reason. Thanks to neuro-science and brain imaging, we know now that the human being is not a rational decision-maker and that emotion is a fundamental partner in human cognition, creativity, and decision-making [37]. The emotional state of humans can be obtained from a wide range of behavioral cues and signals that are available through visual, auditory, and physiological expressions of emotion:

- Emotional state through visual expression is evaluated according to the modulation of facial expressions, gestures, postures, and more generally body language. Data is captured by a camera, allowing non-intrusive configurations.
- Emotional state through auditory expression can be estimated as a modulation of the speech signal [97]. In this case, data is picked up by a microphone, which allows non-intrusive system configurations. In addition, the processing is difficult to handle when more than one voice is present in the audio stream.
- Emotional state through physiological representation is estimated by the modulation of the activity of autonomic nervous system (ANS) [31]. The main limitation is related to the intrusion of detection devices. Also, it is difficult for users to freely manipulate the physiological sensors in relation to facial expressions, body gestures or voices.

Since emotional content reflects human behavior, automatic emotion recognition is a topic of growing interest. Emotions play an implicit role in the communication process compared to the explicit message given by the lexical level. The behavior to be recognized is complex and subtle, presenting diverse manifestations and depending on many factors (social and cultural context, personality of the speaker, etc.).

Emotion Artificial Intelligence Programming, also called Affective Computing (AC), is a key research topic in Artificial Intelligence (AI) dealing with emotions and machines. AC is related to the study and development of systems and devices that can recognize, interpret, process, and simulate human emotions. It consists of the estimation and measurement of human emotions from a variety of data gathered from different sources like speech rate, voice tone, facial expressions, and body gestures.
This thesis explores the use of visual expressions (body gestures) and auditory expressions (speech rate and voice tone) to develop an emotion recognition system. These expressions could be real indicators of the emotional state and significant sources of information when other channels like facial expressions are hidden by keeping a fake smile or neutral face, or there exists a subtle nuance of expressions. There exists a limitless range of applications like:

- E-learning: Tutor expands explanation when the user is in state of confusion (surprised, bored, puzzled) and adds information when he/she is in state of curiosity (happiness, comfort, relaxation).
- E-therapy: Psychological health services (evaluate the psychological state through the patient's gesture).
- Enhanced websites customization: Evaluate a product based on the web surfers' emotions.
- Entertainment: Affect feedback on players' satisfaction to change some characters in the video game.
- Software engineering: Emotions have significant impact on the developers' productivity and code quality. Stress and boredom associated with time pressure will be induced.
- Job interview: The candidate's behaviors during interview session affect the hiring decision. There exists a big relationship between behaviors, interview outcomes, and job performance.

1.2 Non-verbal behaviors

Non-verbal behavior expresses and reveals human emotions and represents, according to social psychologists, 93 % of our interactions with others [8] as shown

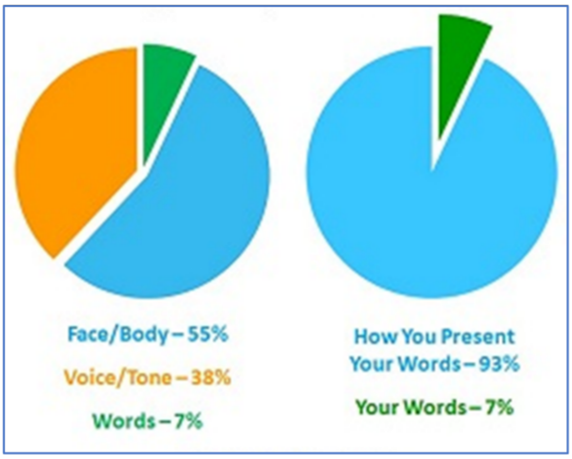

Figure 1.1: Elements of personal communication

in Figure 1.1. Non-verbal behavior can be divided into four categories [3]: proxemics, haptics, kinesics, and vocalics as shown in Figure 1.2.

- Proxemics: relates to how a person uses the space around his/her body in human interactions.
- Vocalics or paralanguages: refers to how speakers express their emotions through voice [154].
- Haptics: refers to the way how a human communicates and interacts through using touching sense. Some self-touch behaviors are related to negative feelings such as stress, psychological discomfort, and anxiety.
- Kinesics: relates to the movement of entire body or body parts such as gestures [42, 65] and facial expressions [169]. When we have spontaneous, unconscious, and non-communicative body gestures, we talk about adaptors as specified in the works of Kipp [95], Ekman and Friesen [39] and McNeill [32].

1.3 Social psychology and job interview

Companies always think about having the best employees. The selection of the most suitable person for the job depends on the answers of the candidate to the interviewer's questions and his/her behavior during the interview session. The study of behaviors is based on the psychological interpretation of the movements and gestures accompanying the answers and discussions. In this context, social psychologists have long studied job interviews in order to understand the relationships between behavior, interview outcomes, and job performance. Several companies give importance to psycho-test based on the observation of the can-

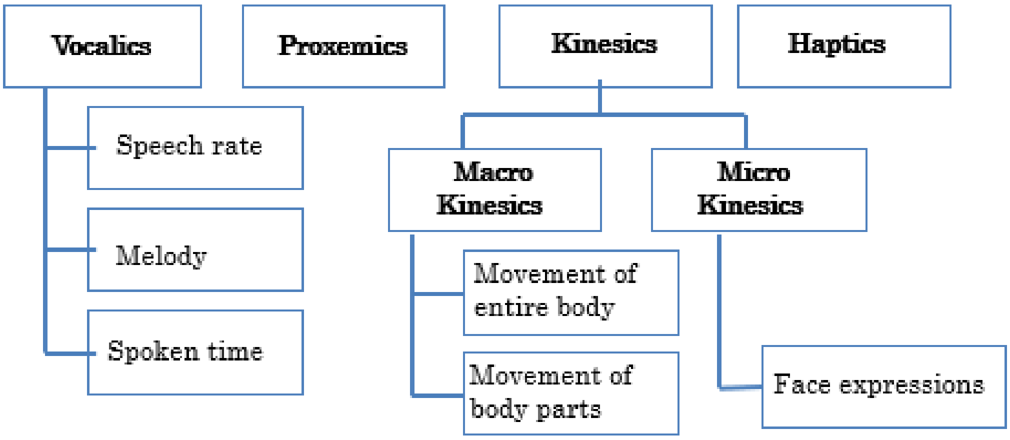

Figure 1.2: Non-verbal behavior categories

didate's behavior more than the answers given especially in sensitive positions like trade, marketing, investigation, etc. Psychology studies used to rely on the utilization of manual annotations by observers but in the last decade, the advent of inexpensive audio and video sensors in conjunction with improved perceptual processing techniques has empowered the automatic and accurate extraction of behavioral cues, encouraging the conduct of social psychology studies. The use of automatically extracted non-verbal cues in combination with machine learning techniques has led to the development of computational methods for the automatic prediction of emotional state. Many works have investigated the reliability [15] (the agreement level between judges for rating candidates) and validity [127] (the amount of relationship between interview ratings and performance) of job interviews, as well as the correlation between high-level social variables like dominance [35], interest [23], emergent leadership [33], personality traits [43], and job performance. Particular attention has been put on the impact of the candidate's non-verbal behavior on the interview outcome. Imada and Hakel [17] affirmed that the candidates who use more non-verbal behaviors (facial expressions, body orientation toward interviewer, eye contact) were perceived as being more competent, motivated, and hirable than candidates who did not. Forbes and Jackson [128] showed that candidates who were recruited nodded more, made more eye contact and hand gesture during interview sessions. The same for Anderson and Shackleton [106] who reported that the most selected applicants produced more facial expressions and gestures during job interview than non-accepted candidates. Parsons and Liden [30] showed that speech characteristics (voice tone, intensity pause, speaking rate, etc.) explained a remarkable amount of variance in the hiring decision. One explanation for the positive connection between candidate non-verbal behavior and recruiting decision can be based on the hypothesis, which establishes that the candidate reveals through his/her immediacy behav-

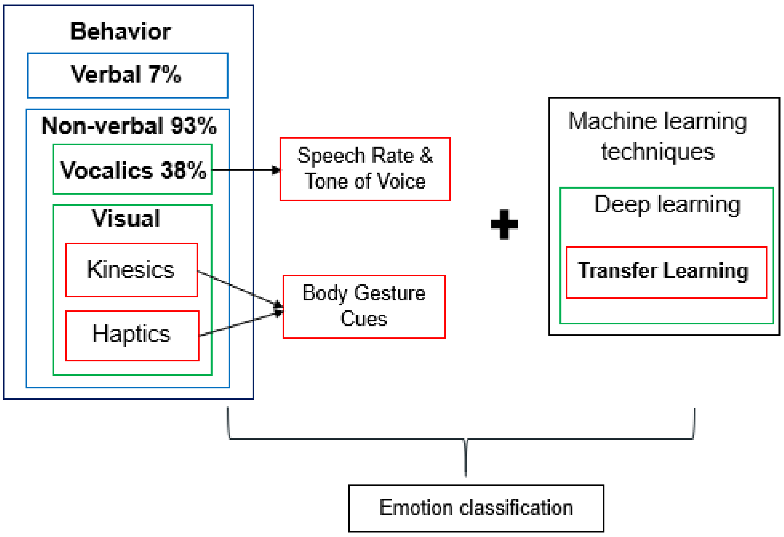

Figure 1.3: From behavior to emotion classification

ior (eye contact, smiling, hand gestures, etc.) a greater perceptual availability, which leads to a constructive outcome on the interviewer and therefore to a positive evaluation.
In this context, these are some applications that could be developed:

- Virtual agents for social coaching
- Job interview simulator
- Hirability impressions in video resumes
- Online selecting service based on pre-recorded questions

1.4 Thesis objective

Our work represents a combination between two interesting research topics in the last decades which are social psychology and affective computing. This combination is realized in order to replace the manual coding performed by an observer with a psychology recognition system based on automatic analysis of non-verbal behaviors by exploiting the strong points of deep learning and transfer learning techniques as shown in Figure 1.3. According to reviews done in the last years, 95 % of researchers focused on facial expressions for emotions analysis [42, 22] and neglected body language or paralanguage (speech) that could be real indicators of the emotional state and significant sources of information when other channels like facial expressions are hidden or when there is a subtle nuance of expressions.

1.5 Open Issues

Emotional body gesture recognition is still an open research problem requiring investigation for many reasons. There exists a subtle nuance of expressions that leads to the confusion in the interpretation of gesture, also the fuzzy nature of emotional states and their instability along time entail difficulties so that, many emotions are still hard to detect. Moreover, searching for a robust method for detection and tracking is a challenging task because spontaneous body gesture characterized by its freedom is very different from specific and predefined gesture in front of the camera for the task of remote control. Then, the quality of training data influences the final results. The addition of several conditions degrades the performance and robustness of an approach and makes the use of such applications more limited. For emotional body gesture recognition from video sequence, we should focus on all these steps: human detection, pose estimation (detection and tracking (body parts or skeleton)), features extraction, and classification.
To attain our objective, these are some questions that should be asked:

- **Structured environment:** Are there restrictions on background, lighting, speed of movement of entire body or body parts?
- **User requirements:** Must the user wear anything special (markers, gloves, glasses, etc)?
- **Features extracted:** Which low-level features (hand crafted features) are computed (edge, region, silhouette, etc.) or is an automatic generation of features possible?
- **Representation of time:** How is the temporal aspect of gesture represented and used in recognition?
- **Body gesture models:** which body model could be useful in our task?

Speech emotion recognition (SER) is also a complex task for several reasons. The first issue of all methods proposed in the literature is the selection of the best features that could be powerful to distinguish between various emotions [20]. Also, the existence of different languages, speaking styles, and accents represents a difficulty because they directly modify the extracted features like pitch and energy. Moreover, we can have more than one emotion in the same speech signal. Each part represents a specific emotion so that defining the boundaries between these parts is a very challenging task. For SER, different steps should be undertaken which are pre-processing of the speech signal, features extraction, and classification.

1.6 Thesis plan

This document presents the released work to develop a psychology recognition system based on automatic analysis of non-verbal behaviors. In this chapter a general overview of the thesis is presented. Motivations and open issues of this work are discussed. Thesis structure as well as the contributions and the list of related publications are presented. The second chapter presents a literature review of previous studies dealing with emotion recognition system. The third chapter is reserved to explain our approach for emotional body gesture recognition using the part-based model. In the fourth chapter, we explain our proposed deep multi-stage approach using the kinematic-based model. The fifth chapter is devoted to present our proposed approach for speech emotion recognition. Finally, we conclude with a general conclusion and possible perspectives.

1.7 Contributions and publications

In this thesis, we propose novel techniques for psychology recognition based on automatic analysis of non-verbal behaviors and precisely the kinesics (body gestures) and vocalics (speech) by exploiting the strong points of deep learning and transfer learning techniques.
The contributions of our work can be summarized as follows:

1. Emotional body gesture recognition is a challenging task because of the complexity of gestures that are rich in varieties caused by high level of liberty. In general, there is no exact definition for the output spaces and mostly based on geometrical representations which could be shallow. In this thesis we try to find the best way to have a good representation of motion information of body gestures that could be useful for the task of emotional classification. In this thesis two models are proposed for emotional body gesture recognition:
    a. Part-based model considers the body as a set of components (hands, head, shoulders, torso, arms) which could be detected separately.
        - Hands have been widely used in comparison to other body parts for gesturing to express the feelings and notify the thoughts. Hands gesture recognition confronts many challenges. In this thesis we propose a hybrid approach for the good representation of hand's local movement in a global temporal template. Our multiple hand detection and tracking method could be useful for Human Computer Interaction applications based on hand gestures.
        - Taking into consideration the coordination between body parts (face and hands), we propose a deep spatio-temporal approach that merges

the temporal normalization method with deep learning method. We demonstrate that deep learning techniques outperform traditional machine learning techniques.

b. Kinematic-based model was used to precise exactly the position of joints in the body and this leads to the good detection and tracking of human skeleton. Human pose estimation is generally used as a separate task in the context of activity and action recognition. In this thesis, we combine pose estimation and emotion classification to propose a new deep multi-stage architecture able to deal with both tasks by exploiting the strong points of models pre-trained. We demonstrate that transfer learning techniques outperform traditional machine learning techniques.

2. Speech is a significant source of information for emotion recognition especially when other channels like face or body are hidden. The shape of the vocal tract, tone of the voice, pitch, and other characteristics are influenced by human emotions. In this thesis, we propose a deep temporal-cepstrum representation of features that is effective in encoding those characteristics of speech. The results obtained prove the effectiveness of our method over existing methods in the state of the art.

Publications The research efforts presented in this dissertation are the summary of these papers:

- Khalifa, I., Ejbali, R., and Zaied, M. (2018). Hand motion modeling for psychology analysis in job interview using Optical Flow-History Motion Image (OF-HMI). The 10th International Conference on Machine Vision, Vienne, Austria.
- Khalifa, I., Ejbali, R., and Zaied, M. (2019). Body gesture modeling for psychology analysis in job interview based on deep Spatio-temporal approach. Parallel and Distributed Computing, Applications and Technologies. Communications in Computer and Information Science, vol 931. pp. 274-284, Springer, Singapore.
- Khalifa, I., Ejbali, R., Schettini, R., and Zaied, M. (2020). Deep Multi-stage approach for emotional body gesture recognition in job interview. The Computer Journal (accepted for publication).
- Khalifa, I., Ejbali, R., Napoletano, P., Schettini, R., and Zaied, M. Deep Temporal-Cepstrum Representation for Speech Emotion Recognition. (future submission to journal).

Chapter 2

Emotion recognition system: Background and related work

2.1 Introduction

Various works have been carried out to develop systems for emotion recognition based on automatic analysis of non-verbal behaviors. Until recently, body gesture and speech have been ignored by the community as a source of emotional information in comparison to facial expressions [42]. In this chapter, we present some notions about emotions like their types and their different models. Based on these, we present the state of the art for automatic emotion recognition using body gesture and speech.

2.2 Notions about emotions

The term emotion is a combination of energy and motion. It is a response of the organism to a particular stimulus (person, situation, or event). It is typically an intense, short-term experience and the person is usually well aware of it. Emotion episodes include various components like action preparation, appraisal of events, expressive behaviors, subjective feelings, and physiological responses [79].

2.2.1 Emotion types

Three types of emotions exist which are: primary or basic emotions, secondary emotions, and social emotions.

a. **Basic emotions:** They are activated by particular events or they manifest in precise circumstances by provoking specific behaviors. They are the basis of our reactions, which are not only determined by our rational judgment or our individual past, but also by our ancestral past. There are six primary emotions: joy, sadness, anger, fear, disgust, and surprise. In fact, these basic emotions are like a primary material, from which all other emotions can be made [37].

b. **Secondary emotions:** They increase the intensity of reactions over time [37]. Some of these are triggered by thinking about what might have happened or not, unlike basic emotions which are triggered by actual occurrences (direct reaction to external event). When we feel angry, we may feel ashamed afterward or when we feel happy, we may feel proud. Several secondary emotions exist like pride, trust, confidence, relaxation, disappointment, boredom, uncertainty, anxiety, confusion, etc.

c. **Social emotions:** These emotions are inherent in the relationship with others like guilty, jealousy, timidity, humiliation, etc. All these emotions are learned and are built up from the primary emotions [37].

2.2.2 Emotion models

The manipulation of emotions with computer raises many issues. First, at the level of their representations, it is a question of finding a formalism that agrees with the existing psychological results, while allowing a simple manipulation. Then, for a given event, it is necessary to be able to determine the emotional potential associated with it. Based on the work in social psychology [5], some measures consider emotional states as categories, others as a multidimensional construct.

2.2.2.1 Categorical model

It is the most popular model in the literature. The universal character of emotions leads to the definition of basic emotions that can be observed in all individuals regardless of their ethnicity or culture. Based on the research of Ekman [112, 113], this model divides emotions into a set of distinct classes that can be described easily. Therefore, the affective denominations that don't find their place in these classifications are considered as a mixture of primary emotions. The rationale behind the use of this model is that these basic emotions are clearly identifiable in the majority of individuals, especially through non-verbal communication. However, the number of emotions that should be considered is still an open question because this number differs from a researcher to another as shown in Table 2.1.

Author	Emotion classes
Ekman [114]	anger, disgust, fear, joy, sadness, surprise
Tomkins [148]	anger, interest, contempt, disgust, distress, fear, joy, shame, surprise
Izard [29]	anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise
Plutchik [124]	acceptance, anger, anticipation, disgust, fear, joy, sadness, surprise

Table 2.1: Classification based on categorical model

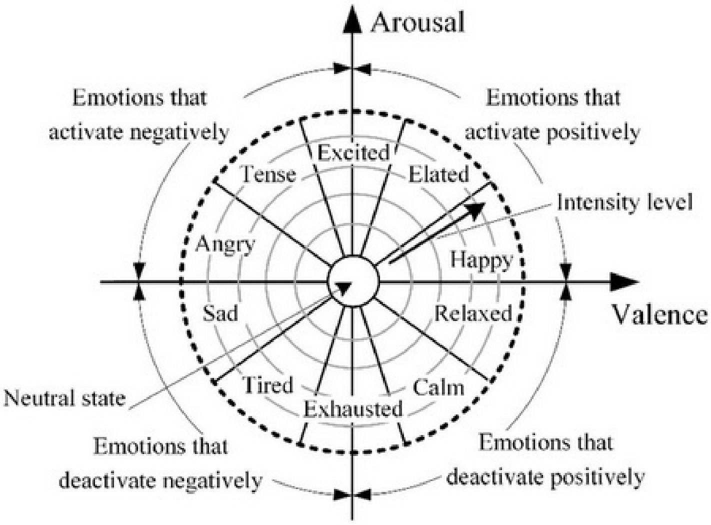
Arousal
Emotions that activate negatively
Emotions that activate positively
Excited
Tense
Elated
Intensity level
Angry
Happy
Valence
Sad
Relaxed
Tired
Calm
Exhausted
Neutral state
Emotions that deactivate negatively
Emotions that deactivate positively

Figure 2.1: Dimensional representation of emotions: Russel's model [36]

2.2.2.2 Dimensional model

It is a popular and theoretical approach in the psychology of human emotions [68, 92], which proposes a continuous representation on several axes or dimensions. The dimensions include valence (positive or negative character of the emotional experience), arousal or activation (how a person acts under the emotional state) and control (how to control over emotion). Many automatic emotion recognition systems are based on the dimensional representation like the Russell's model [69] due to its simplicity (it divides the space into a limited set of categories like positive vs negative) as shown in Figure 2.1.

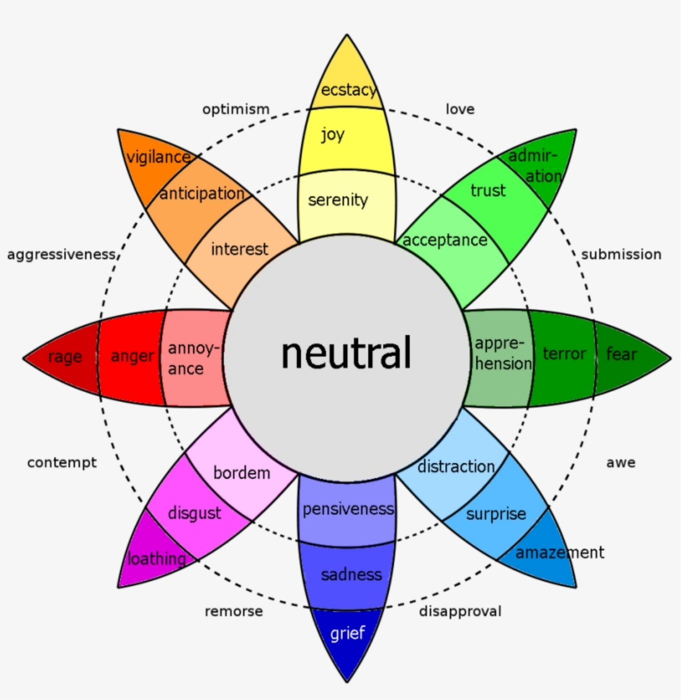

Figure 2.2: Componential representation of emotions: Plutchik's model [42]

2.2.2.3 Componential model

It is in between the dimensional and categorical models; it arranges the emotions in a hierarchical way where the superior layers are composed of emotions from previous layers. According to the Plutchik model [125] as shown in Figure 2.2, complex emotions are the combination of pairs of primary emotions like: Disapproval = Surprise + Sadness and Contempt = Disgust + Anger, etc.

2.3 Emotional body gesture recognition: State of the art

2.3.1 Body gesture cues

Gestures are a crucial component in the interpersonal communication as they are used to decode the vocal content and aid observer comprehension by activating the images in the listener's mind and reinforcing the attention. Gestures can be identified either from an interpersonal interaction (non-verbal, semiotic communication) or from a physiological signal (reflex result or voluntary muscle contractions). We can classify the gestures according to the body parts involved [131]: Gestures involving the whole body, head posture and facial expressions, and gestures of the hands that form the main category of interactive gestures. The

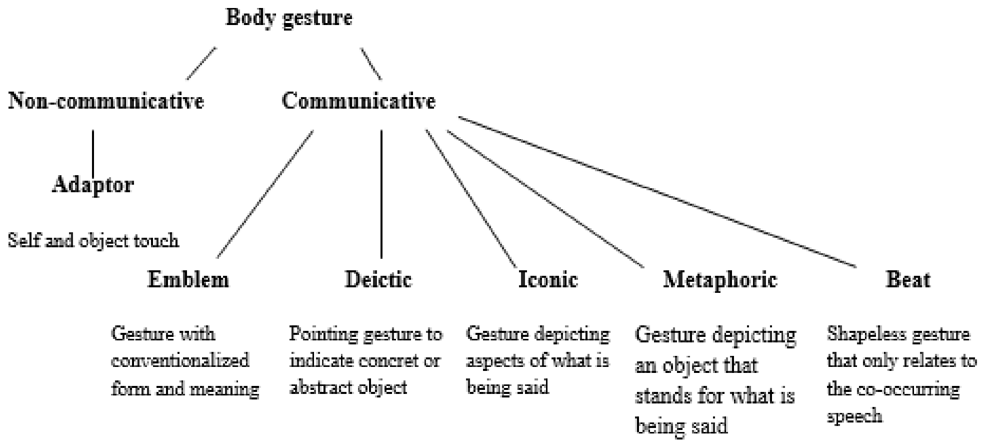
Body gesture
Non-communicative
Communicative
Adaptor
Self and object touch
Emblem
Deictic
Iconic
Metaphoric
Beat
Gesture with conventionalized form and meaning
Pointing gesture to indicate concret or abstract object
Gesture depicting aspects of what is being said
Gesture depicting an object that stands for what is being said
Shapeless gesture that only relates to the co-occurring speech

Figure 2.3: Gesture categories [95]

research in this area is linked to the recognition of hand positions [60, 152, 64], the development of human-machine interactions, and the interpretation of sign language. As shown in Figure 2.3, several researchers proposed various gesture classifications; Kipp [95] used a set of six classes of gestures detailed in the work of Ekman and Friesen [39] and McNeill [32] which are: emblem, deictic, iconic, metaphoric, beat, and the last one is adaptor that represents the non-communicative and spontaneous gesture. It is usually unintentional and includes self-touch behaviors such as touching an object or the own body (scratching head, touching nose, etc). It can inform about the state of the speaker and it generally reflects negative emotions like anxiety and uncertainty. Body has been used for gesturing to notify the thoughts and express the feelings. Based on previous works of researchers like Ekman and Friesen [39] and Gelder [21], Witkower and Tracy [169] wrote a review about body expressions of emotions like pride, shame, disgust and embarrassment and they presented the existing coding systems for emotional analysis. In another work, Noroozi et al. wrote a survey on body gestures and their interpretations [42]. Different emotion classifications are presented: Baltrusaitis et al. [151] proposed Geneva Multi-modal Emotion Portrayals-Facial Expression Recognition and Analysis (GEMEP-FERA) database that consists of 10 actors displaying 5 emotions including anger, fear, relief, sadness, and joy. Baveye et al. [160] created a video database for affective content analysis (LIRIS-ACCEDE), which contains upper bodies of 64 actors, extracted from different kinds of movies like action, drama, and romance, displaying 4 emotions (sadness, anger, disgust, and fear). As a result of the feedback obtained from the works of Ekman and Friesen [39], Burgoon [74] and Coulson [90], Gunes and Piccardi [56] identified the correlation between body gestures and the emotional state categories as presented in Table 2.2. They were not limited to the six basic emotions

Emotion	**Body gesture**
Happiness	Arms opened, hand clapping, hands made into fists and kept high.
Sadness	Covering face with hands, trunk leaning forward, dropped shoulders, body extended and hands over the head.
Anger	Hands on waist, lift right and left hand up, finger point with right or left hand.
Disgust	Hands covering the neck, backing, hand on the mouth.
Surprise	Two hands covering the cheeks or the mouth.
Fear	Crossing arms, covering the body parts, arms around the body/shoulders.
Boredom	Hands below the chin, elbow on the table.
Anxiety	Tapping tips of fingers on table, hands pressed together in a moving sequence, biting the nails.
Uncertainty	Palms up and shoulder shrug, right/left hand touching the chin, forehead, nose, ear or the neck, scratching hair or head.

Table 2.2: Body gesture cues and their interpretations

which are happiness, sadness, disgust, surprise, fear, and anger, Gunes and Piccardi added secondary emotions like anxiety, uncertainty, and puzzlement that can be real indicators of the emotional state and affect the decision.

2.3.2 Human body models

Body gesture recognition is a challenging task due to the complexity of gestures that are rich in varieties caused by high level of liberty. Numerous models have been proposed for the representation of human body [12] using part-based multiple patches (part-based model), skeleton (kinematic-based model), centroid, multiple points, rectangular or elliptical patch, contour, and silhouette as shown in Figure 2.4. Part-based model and kinematic-based model as presented in Figure 2.5 are the most robust ways for modeling the human body in automatic processing [42] and especially when we need to identify some body parts that could be useful in the task of emotion recognition.

2.3.2.1 Part-based model

Using this model, the body is considered as a set of components (head, shoulders, hands, arms, torso) which could be detected separately with their specific detec-

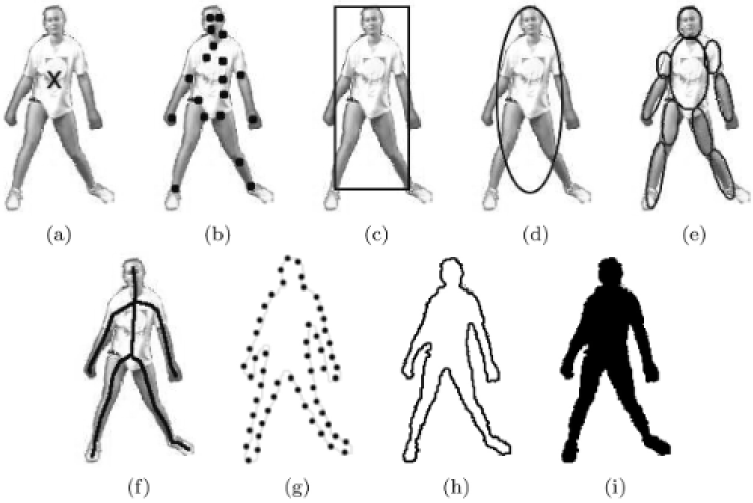

Figure 2.4: Object representations [12]:
(a) centroid, (b) multiple points, (c) rectangular patch, (d) elliptical patch, (e) part-based multiple patches, (f) skeleton, (g) contour, (h) control points, (i) silhouette

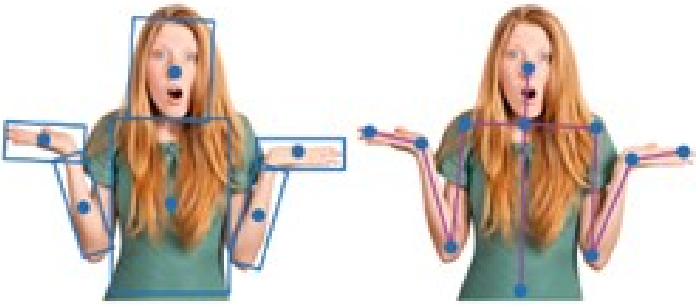

Figure 2.5: Part-based model vs kinematic-based model [42]

tors. According to studies done by social psychologists like Mehrabian [69, 8], hand gestures give clues about the emotional state [12, 61] of the speaker and they can even help a person become a better communicator. We might also consider the relationships between body parts [65] like face and hand that could be effective for identifying emotions.

Taking as example the work of Marcos-Ramiro et al. [7], they suggested a technique to construct Hands Likelihood Map based on Optical Flow to detect and track the upper body parts with more energy which are the two hands. The motion of hands will be then classified into four classes (hand on table, hidden hand, gesturing, and self-touch).

N°	COCO output format	MPII output format	Pose Evaluator output format
0	Nose	Head	Head
1	Neck	Neck	Torso
2	Right Shoulder	Right Shoulder	Right Upper Arm
3	Right Elbow	Right Elbow	Right Lower Arm
4	Right Wrist	Right Wrist	Left Upper Arm
5	Left Shoulder	Left Shoulder	Left Lower Arm
6	Left Elbow	Left Elbow	
7	Left Wrist	Left Wrist	
8	Right Hip	Right Hip	
9	Right Knee	Right Knee	
10	Right Ankle	Right Ankle	
11	Left Hip	Left Hip	
12	Left Knee	Left Knee	
13	Left Ankle	Left Ankle	
14	Right Eye	Chest	
15	Left Eye		
16	Right Ear		
17	Left Ear		

Table 2.3: Keypoints output format for pose estimation

2.3.2.2 Kinematic-based model

To estimate the body pose, kinematic-based model was proposed by some researchers for the tasks of activities and actions recognition [40, 9] and the results prove its efficiency. This model is a collection of interconnected joints as shown in Table 2.3. Based on the kinematic model, some challenging datasets have been proposed in the last few years to make the task of pose estimation less difficult. The most popular ones are:

- Common Object in Context (COCO) dataset with 18 keypoints [10].
- MPII Human Pose dataset with 15 keypoints [103].
- Human Pose Evaluator dataset with 6 parts [107].

2.3.3 Existing techniques for gesture recognition

Gesture recognition has been gradually related to the development of systems able to identify human gestures and decode them to enrich the user experience. It is a complex task implying different aspects like motion analysis, motion modeling, machine learning and even psycho-linguistics studies.

Gesture recognition was utilized in multiple domains like video surveillance, hu-

man computer interaction (HCI), robotics, decision system, etc. Several tools were used in the gesture recognition systems such as image and video processing, pattern recognition, statistical modeling, computer vision, etc. For decades, this topic was studied by several researchers and each one intervened in a specific phase: Human detection, pose estimation (detection and tracking), feature extraction, classification, or regression. In general, the support in which the gesture information is stored is a video. It is then indispensable to apply image and video processing techniques and be able to exploit them efficiently.
Several techniques exist in literature [138, 129]:
- For detection: Skin color, contour, background subtraction, Adaboost method.
- For tracking: Point, kernel, and silhouette tracking.
- For recognition: There exists two categories which are static and dynamic:

- For static recognition: Linear and non-linear classifier.
- For dynamic recognition: Hidden Markov Model (HMM), Dynamic Time Warping, Time Delay Neural Network, Finite State Machine, etc.

2.3.3.1 Hand gesture modeling

In the last decades, there existed a various range of applications based on body gesture recognition. For Human Computer Interaction (HCI) applications, hands have been widely used in comparison to other body parts. Hand detection and tracking in a video sequence confronts many challenges: Complex background when there are other objects in the scene with hand, complex shapes and motion, variation of hand positions in different frames leads to erroneous representation of features, hand poses with different sizes in gesture frame, overlap with other regions in the image, etc. Many researchers have strived to improve the hand gesture recognition techniques. These techniques could be divided into wearable glove-based sensor approach and camera vision-based approach [96, 142].

Hand gestures using Wearable glove-based sensor approach For this technique, several sensors were used such as angular displacement sensor, flex sensor, accelerometer sensor, curvature sensor, etc. They can provide the exact coordinates of palm and finger locations, orientation, and configurations [115].
Several glove systems were proposed by research laboratories over the past 3 decades [83] like Human Glove, Cyber Glove, 5DT Data Glove, and Pinch Glove as shown in Figure 2.6, and only a few of them became commercially available. Taking as example the work of Bedregal et al. [25], who proposed a method for hand gesture recognition based on fuzzy logic, and they applied it for the recognition of Brazilian language gestures. The suggested method used a glove with

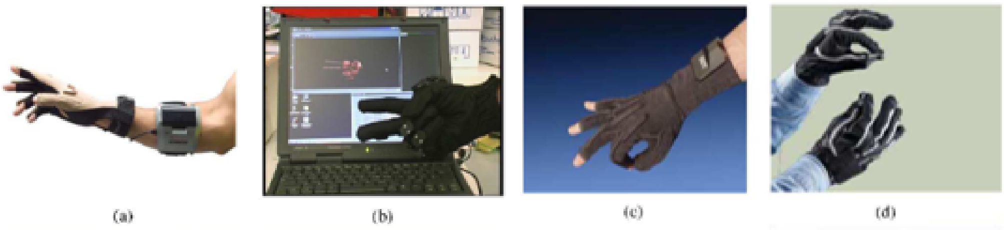

Figure 2.6: Glove systems examples [83]:
(a) Cyber Glove, (b) Human Glove, (c) 5DT Data Glove, (d) Pinch Glove

19 sensors distributed on the segments of the different fingers. The authors supposed that a gesture can be seen as a sequence of frames where each frame image has a very specific configuration of the image segments. The set of finger angles and the distances between the fingers with fuzzy logic were used to recognize the gestures. Mahdi and Khan [141] applied a glove with sensors to extract features from the signs and used an Artificial Neural Network to identify 24 letters of the American Sign Language alphabet.

However, some **limitations** make the wearable glove-based sensor unsuitable for use when people suffer from chronic diseases that result in the loss of muscle function. These sensors may also cause skin damage and infection.

They can also lead to a false interpretation of the emotional state because wearing something special increases the level of stress and makes the user feel discomfort. Moreover, some sensors are quite expensive.

Hand gestures using camera vision-based approach With the evolution of open-source software libraries, many algorithms were developed based on computer vision methods [59]. They are more suitable for use and easier than wearable glove-based sensor methods for the detection of hand gestures. Hand gestures could be exploited under various range of applications such as sign language [75], clinical operations [72], virtual environments [130], home automation [137], decision system in job interview [7], etc.

Based on the works done by Kaur et al. [59], Murthy et al. [52], Khan et al. [80], Suriya et al. [126], and Sonkusare et al. [73], Oudah and Naji [96] wrote a survey on hand gesture recognition. They focused on computer vision techniques where they described seven common methods such as skin color, appearance, motion, skeleton, depth, 3D-module, and deep learning.

Skin color is one of the most popular methods for hand detection and segmentation. Liang et al. [61] used a cascade hand detector with skin detection and Adaboost classifier to isolate the hand. Then, they applied a point-based tracker called Median-flow tracker to estimate the hand motion information in short-term tracking.

The idea presented in [100, 123, 63, 152] is to utilize the skin color for hand detection in YCbCr space, separate the video regions and analyze the areas founded by a neural network in order to determine the region containing the hand. Then, an approach inspired from point tracking approach and K-Nearest Neighbor (KNN) is applied for hand tracking.

Choudhury et al. [2] proposed an approach based on the combination of frame differencing method and skin color segmentation, however this method is still sensitive to scenes that contain moving objects in the background.

Stergiopoulou et al. [38] merged motion-based segmentation with skin color and morphology features to overcome the problem of complex background.

Bergh et al. [102] suggested a hybrid method based on histogram and Gaussian Mixture Model (GMM). Chen et al. [122] proposed an approach for hand recognition based on Haar-like features and Adaboost learning algorithm.

Kulkarni and Lokhande [157] used histogram technique and edge detection like Sobel, Canny and Prewitt operators for segmentation and hand detection, then they applied the feed forward back propagation artificial neural network for classification.

Fang et al. [163] suggested an extended Adaboost for hand detection and Optical Flow (OF) with color cue for tracking.

In another work [162], they offered a fast method for detecting and recognizing hand gestures. First, the approach uses the integral image to approximate the Gaussian derivatives for the calculation of the image convolution to determine the descriptors. Then, multi-scale geometric characteristics are obtained from the points to represent the hand gestures.

Prakash and Gautam [66] used YUV color space with CamShift algorithm for hand detection and Naive Bayes classifier (NB) for gesture recognition.

The work presented in [150] used the method of Kalman Filter (KF) for hand tracking in a video sequence. Then, Canny and Harris corner was used for feature points' extraction.

Chen et al. [168] proposed a method for real-time hand gesture recognition using finger segmentation. A background subtraction method was used for hand region detection. Then, the palm and fingers were segmented. Fingers in the hand image were recognized and a simple rule classifier was used to predict the labels of hand gestures.

Crampton et al. [136] proposed a real-time system called Counter Finger that interprets certain hand gestures as input to the computer. This system used two new techniques: Background differencing method adaptive to changing lighting conditions and camera movement, and a method to analyze the contours of the hand.

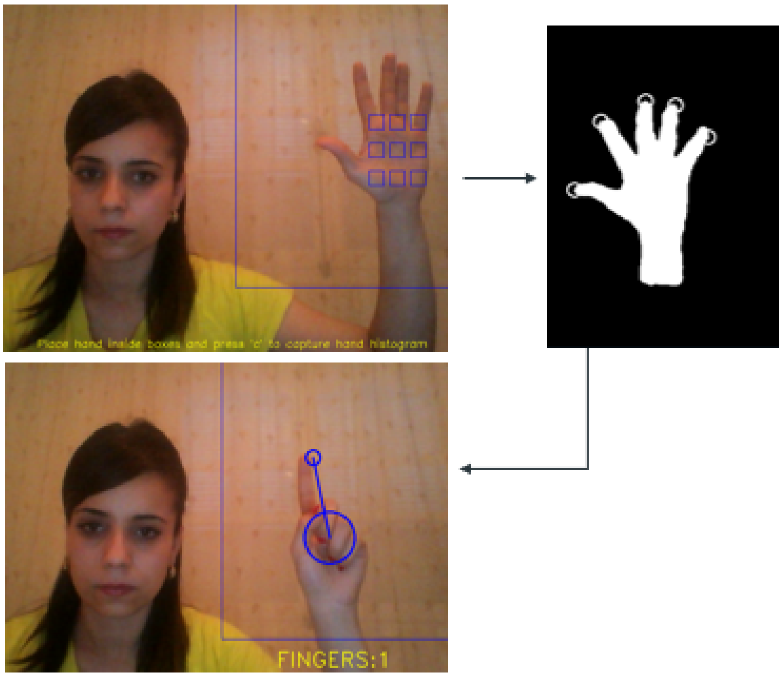
Place hand inside boxes and press 'c' to capture hand histogram
FINGERS:1

Figure 2.7: Real-time hand detection and finger counting

Figure 2.7 presents an example of real-time hand detection and finger counting using Histogram based approach and contour detection. Kinect depth sensor is one of the most popular cameras that was used for gesture recognition in a wide range of applications. Li and Yi [87] applied a technique for hand gesture classification using Microsoft Kinect camera. The system can identify 9 gestures in a predefined gesture scenario. The extraction of the characteristics is based on the identification of fingers.

Chen et al. applied the Kinect camera depth sensor for acquisition and segmentation. Then, the Support Vector Machine (SVM) algorithm and HMM were used for classification [161].

Wang et al. [27] suggested a hand gesture recognition system based on a new super-pixel distance "Earth Movers Distance" used in conjunction with the Kinect depth camera to measure the dissimilarity between hand gestures. This measurement is not only robust to the deformation and articulation, but also invariant to scale, translation, and rotation.

Some researchers proposed approaches for image-based hand keypoint localization. Generating annotated hand keypoints datasets presents a major challenge. In fact, there are no robust markerless hand keypoint detectors that work on RGB images in the wild with fewer cameras and in less controlled environments. Simon et al. [153] proposed a Multiview bootstrapping method that can localize hand joints in RGB images without requiring depth. However, the annotation in single images is difficult because joints are often occluded due to articulations of other parts of the hand or a particular viewing angle. This method fails and the average 2D error in pixels increases when there is another object with hand or when the two hands are present in the scene. So, the main objective for many researchers in this field is how to create a robust hand detector that more closely reflects real world capture conditions.

2.3.3.2 Body gesture modeling

Body gesture recognition is a challenging task due to the complexity of gestures that are rich in varieties caused by a high level of liberty. Several techniques were proposed for the representation of human body. However, there is no exact definition for the output spaces and mostly based on geometrical representations which could be shallow. For that reason, each researcher tried to find the best way for the good representation of body motion information that could be useful for the task of emotional classification.

Taking into consideration the relationships between body parts, different methods were proposed for emotional body gesture recognition.

Gunes and Piccardi [58] applied a maximum voting of apex frames for key frame extraction. Then, the features were extracted based on OF, comparison to the neutral frame, edge, and geometry, and for classification they applied SVM, HMM and Random Forest (RF).

In [57], they used a combination between silhouette based model and color for upper body location (head, torso and hand), background subtraction for segmentation and then the CamShift technique for multiple hand tracking with comparison of bounding rectangles to predict the location of hand in the next frame. For the recognition task they used a BayesNet classification algorithm.

Chen et al. [135] proposed an approach based on the dynamics of expression using a complete cycle (neutral, onset, offset and apex phases) instead of using just the candidate's key frames for each video. For feature extraction, they combined the Histogram of Oriented Gradients and Motion History Image (MHI-HOG) and Image-HOG through Bag of Words model (BOW) and Temporal Normalization algorithm (TN) to obtain a representation of features that will be the input of SVM classifier.

Castellano et al. [44] extracted some features which are the contraction index and quantity of motion of upper body, also the acceleration, fluidity and velocity of head and limbs. They used Bayesian Network (BN) classifier to distinguish between 4 emotions: Anger, joy, sadness, and pleasure and achieved 61 % as recognition rate. In another work [45], they compared between different classifiers like Decision Tree (DT), Hidden Naive Bayes (HNB), Dynamic Time Warping with

1-Nearest Neighbor (DTW-1NN) and they concluded that the last one gave the best results to classify emotions based on body gestures.

Saha et al. [145] applied skeletal geometrical features as inputs for some classifiers like Ensemble Tree (ET), Binary DT, KNN and SVM. The best results obtained using ET to classify five emotions: fear, anger, sadness, happiness, and relaxation.

Kapur et al. [4] used a 3D motion capture system to record the archetypal body movements. They tested some classifiers to recognize four emotional states and they concluded with the effectiveness of SVM and Neural Network (NN) that were able to achieve 84 % of recognition rate.

Piana et al. [144] proposed a method to real-time emotional body gesture recognition. The descriptors were extracted from 3D motion clips containing full-body motions which are recorded using Microsoft Kinect and optical motion capture system. The joints of the body were tracked, and the motions feature vectors were extracted. These vectors will be the input of linear SVM to classify six emotions. However, human validation demonstrated that three of the emotions (fear, disgust, and surprise) were confused with each other. In fact, this approach proved its effectiveness when classifying just 4 emotions (Sadness, happiness, anger, and fear).

In the last years, deep Convolutional Neural Networks (CNN) [6] have been proposed for large-scale image processing. So, researchers moved from methods based on hand crafted and low-level features (edge, texture, color, etc.) which are the input of a classifier to deep learning technique and all its variants.

Barros et al. [110] proposed a Multi-Channel Convolutional Neural Network approach (MCCNN) that was applied sequentially to have a hierarchical feature representation of 3 channels which are grayscale image, Sobel in direction x and Sobel in direction y by applying an edge detector which is the Sobel Filter. These three channels represent the low-level features that will be the input of the CNN with two layers.

In another work, Barros et al. [111] implemented a Cross-Channel Convolution Neural Network (CCCNN) model to learn the location of expressions in a cluttered scene. They applied the probability distributions to estimate the location of interest for each sequence of frames by its position on the x and y-axis using the filtering capability of the convolution layers to learn.

Thai Ly et al. [147] extracted the key frames from video, then Recurrent Neural Network called "Convolutional Long Short-Term Memory (LSTM)" was used for exploiting the sequence information. Firstly, they extracted the features from video frames using CNN to have feature maps that will be after that binarized using Iterative Quantization. Secondly, they applied the Hamming distance between consecutive frames to retrieve the key frames. Then, they used the HMI to overlap the key frame sequences in one image that will be the input for LSTM.

Human pose estimation has largely focused on finding the body parts of individuals. It has been used generally for actions and activities recognition [40, 54]. Some researchers tried to refine the heat maps to improve the capacity of learning human skeleton structures. Cao et al. [167] developed the first open-source real-time system for multi-person 2D pose detection called "Open Pose". Their proposed method takes the entire image as the input for a CNN to predict confidence maps for body part detection and Part Affinity Fields for part association. The parsing step performs a set of bipartite matchings to associate body part candidates. Then, they assembled them into full body poses for all people in the image. This technique achieved high accuracy and real-time performance. Kolotouros et al. [108] proposed a self-improving method for training a deep network for 3D human pose and shape estimation through a collaboration between regression method and optimization-based method. Their proposed approach uses the network to provide an initial estimate to the optimization routine, which then fits the model in the loop and provides model-based supervision for the training of the network.
These works were the source of inspiration to develop our deep multi-stage architecture able to deal with both tasks: Pose estimation and emotion classification for emotional body gesture recognition. Several applications were developed based on automatic recognition of emotional body gesture [42] like videoconferencing, violence detection, video surveillance, psychological research tools, job interview simulator, etc.
In fact, body gesture conveys rich emotion but usually requires a camera pointing at the subject or wearable sensors which are not suitable for spontaneous emotion and thus for real world applications (e-learning, e-health, etc.). Real world applications need widely available and economics camera that can be placed in the environment with minimum effort such as smartphones, personal camera, etc. Hence, recognizing emotion using 2D joints extracted from simple RGB video inputs is highly desirable in this context. In this thesis, to deal with spontaneous, free, and non-constrained body gestures for human-human communication, we propose efficient deep Spatio-Temporal body motion analysis and 2D RGB video-based human pose detection methods. The proposed approaches are interesting for job interviews because AI is already replacing parts of the interview process. Moreover, most of the works on emotion recognition in the literature focus on classic and basic emotions such as (happy, sad, disgust, etc.). The current work addresses under studied emotions such as (anxiety, uncertainty, puzzlement, etc.) which are challenging to detect and to recognize.

Emotion	Vocalic cues
Neutral	Speech is articulated clearly with adequate pauses between words.
Happy	Raised precision of articulation, more breathy sounding voice. Pitch mean, pitch range and pitch variance increase.
Sad	Overall reduction in articulation, lower than normal average pitch. Pitch range is narrow with a slow tempo.
Angry	Highest energy level and pitch, faster rate of speech.
Disgust	Increased precision of articulating and stressing certain words, descending pitch inflection at the end of words. Rate of speech slow and with many pauses, longer phonation time.
Boredom	Hands below the chin, elbow on the table.
Anxiety	Tapping tips of fingers on table, hands pressed together in a moving sequence, biting the nails.
Surprised	Pitch median and tempo are high with quite a wide pitch range.

Table 2.4: Vocalic cues and their interpretations [53]

2.4 Speech emotion recognition (SER): State of the art

2.4.1 Vocalic cues

SER is an interesting research topic in the fields of behavioural sciences and affective computing dealing with machines and psychology. Human speech is a significant source of information especially when other channels, like face or body, are hidden. It is influenced by the physiology of the speaker, shape of vocal tract, tone of the voice and the phonetic content such as pitch, intensity, energy, and duration [154].

Since these characteristics are influenced by emotions, the selection of the right speech descriptors is fundamental to achieve an automatic discrimination and recognition of distinct emotions like neutrality, happiness, sadness, anger, disgust, and surprise as shown in Table 2.4[53].

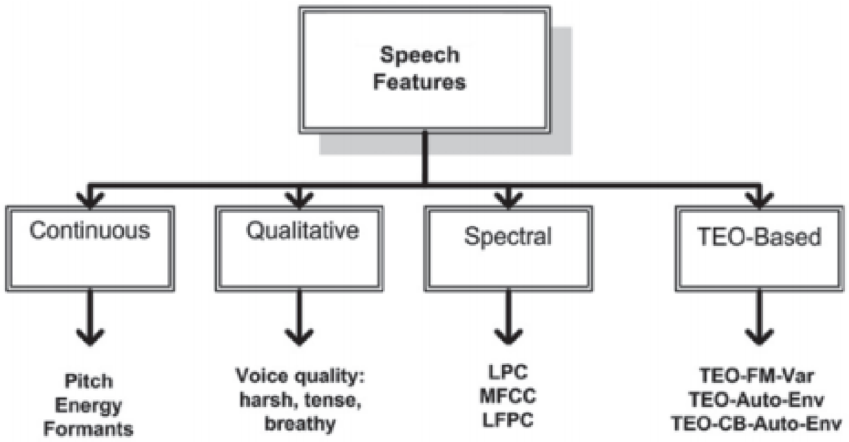

Figure 2.8: Acoustic features categories [91]

2.4.2 Acoustic features for SER

Following the categorization proposed by Ayadi et al. [91], the most used features for SER are grouped into four categories [155] as shown in Figure 2.8 : (1) Continuous - such as pitch, energy, and formants; (2) Qualitative - such as voice quality, harshness, breathy; (3) Spectral - such as Linear Predictive Coding (LPC), Mel-Frequency Cepstral Coefficients (MFCC); (4) Teager Energy Operator-based features (TEO) such as TEO-FMVar, TEO-CB-Auto-Env.
Singh et al. [143] provided a comparative study between two existing techniques that are available for feature extraction based on some parameters which are vocal tract, features type and set, secure, filters, speaker verification, channel effect, performance, and speech sample as shown in Table 2.5: Cepstral features equivalent to short-term features (i.e. MFCC) and non-cepstral features called also long-term features (prosodic). In SER, features have been experimented in combination with several classification methods, such as: HMM [14], SVM [49], GMM [156], NN [116], brain emotional learning model (BEL) [101], Voiced Segment Selection (VSS) [48], and very recently deep learning architectures [19].

2.4.3 Existing techniques for SER

In the last decade, different experiments were carried out to identify emotions from speech from different accents and languages. Many researchers focused on identifying the best features and model architecture for SER.
Rong et al. [67] proposed Ensemble RF to Trees method with a high number of features like spectral-related, contour-related, tone-based, vowel-related features, etc. They demonstrated that this method performed better than other dimension

Parameter	**MFCC**	**Prosodic**
Vocal tract	Depends on shape of the vocal tract	Excitation of the vocal tract and the speaking style
Features set	Uses a small set of standard features	Uses long term features
Features type	Uses cepstral features	Uses non-cepstral features
Secure	Not easy to mimic	Easier to mimic
Filters	Uses filter bank	Does not use filters
Speaker verification/identification	It gives better results for both	It gives better result for speaker verification
Performance	MFCC lonely able to perform well	Prosodic features alone cannot perform well
Channel effect	Cepstral features affected by the channel distortion	It is believed that prosodic features are less vulnerable to the channel distortion
Speech sample	It requires less speech sample, less time and not so computationally complex	It requires a lot of speech samples, time consuming and computationally complex

Table 2.5: A comparative chart: cepstral and non-cepstral features [143]

reduction methods like Multi-Dimensional Scaling (MDS) and Principal Component Analysis (PCA). For classification they used RF and DT.

Wu et al. [149] used Semantic Labels (SLs) for feature extraction as a first experiment then, they proposed a fusion method based on the combination of SLs and acoustic-prosodic features. They proved its effectiveness with GMM, SVM, and Meta DT classifiers.

Yeh et al. [71] applied Segment-based method for SER in Mandarin speech. This approach is based on the Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) for feature extraction and they used the KNN (K sets to 10) as classifier.

Dai et al. [158] proposed a computational approach for SER: They extracted 25 acoustic features like Short-time Energy, Pitch, Zero, Crossing Rate, first and second Formant from speech signals collected from voiced social media like Wechat. Then they applied Least Squares-Support Vector Regression (LV-SVR) model for emotion classification.

Grimm et al. [93] extracted acoustic features like pitch, energy, speaking rate and spectral characteristics. They were the inputs to a multi-dimensional model with three dimensions (valence, activation, and dominance), these dimensions were mapped to some emotion categories using KNN classifier.

Lee et al. [28] used hierarchical structure for binary DT. This method used

large-margin features that were the input of binary classifiers (Bayesian Logistic Regression and SVM).
Tanmoy et al. [155] used features based on Discrete Wavelet Transform (DWT) to decompose the speech signal combined with three methods: Support Vector Classifier (SVC), Gaussian Naive Bayes (GNB) and KNN.
Lampropoulos and Tsihrintzis [86] used MPEG-7- low level audio descriptors (Basic spectral and Timbral) combined with a Radial Basis Function SVM (RBF-SVM).
Zamil et al. [13] used static 13-dimensional MFCC features as input to the Logistic Model Tree (LMT) classifier with linear regression. Each frame in the utterance was classified and a vote was assigned to that emotion class and then they applied the majority voting mechanism: The emotion with maximum votes will be considered to classify all the input signal.

Some researchers applied a combination of different acoustic features. According to them, extracting more features leads to a good classification accuracy. However, the increase of the vector features dimension may lead to the redundant of information and affect negatively the classification time and accuracy [91].
Shegokar and Sircar [117] selected features based on continuous wavelet transform (CWT) and 278 prosodic features combined like linear predictive coding, zero crossing rate and entropy. For classification they applied SVM with a Quadratic kernel (Q-SVM).
Chen et al. [82] applied a combination of some features: Zero Crossing Rate (ZCR), pitch, formants, spectrum centroid, spectrum cut-off frequency, correlation density, fractal dimension, and five Mel-frequency bands energy. Then, they used Fisher and PCA to reduce the dimension of vector features. This vector was the input of SVM and Artificial NN classifiers.
In [170], they extracted a mixing feature set including speaker independent features like the fundamental frequency and dependent features such as rhythm characteristics, then they used feature selection method to eliminate the redundancy based on the combination of correlation analysis and Fisher. These features were the input of the DT for the task of emotion classification.
In the work of Deshmukh et al. [46] prosodic and spectral features were used: Short Term Energy (STE), Pitch and MFCC coefficients. These features are processed to multiclass SVM to classify 3 emotions.

Some researchers applied deep learning techniques like Parry et al. [119], who used spectrograms as input of a combination of both a CNN and LSTM. Badshah et al. [19] used a spectrogram as input of a model similar to AlexNet with a rectangular kernels of varying shapes and sizes, along with max pooling in rectangular neighborhoods. Popova et al. [120] also used spectrograms as input of a pre-trained VGG-16.

Guo et al. [164] decomposed the speech using Dual-Tree Complex Wavelet Transform, they reconstructed the noise-like interference for speech augmentation and then they extracted the acoustic features using SliCQ-Non Stationary Gabor Transformation (SliCQ-NSGT) and ComParE feature set from OpenSmile Toolkit that will be the input of SVM and Bi-LSTM.

In the last decade, SER was largely employed in several application domains like human computer interface, robotics, audio surveillance, e-learning, computer games, decisional system for job interview, etc [19].

2.5 Discussion

In literature, there exists different modalities for psychology recognition like facial expressions, body language, speech, physiological signals. Emotional Body gesture recognition is a challenging task due to the complexity of gestures that are rich in varieties caused by high level of liberty. Several techniques were proposed for the representation of human body. Hand has been widely used in comparison to other body parts and many applications were developed based on hand gestures as inputs. Hand detection and tracking confront many problems like complex shapes and motions, rotation in any directions, background problem, partial or full occlusions, etc.

To overcome these problems, some approaches were proposed. They are divided into wearable glove-based sensor approach and camera vision-based approach [96, 142]. The use of wearable sensors could be suitable for specific and precise hand gestures for computer monitoring for example. However, it is unsuitable for spontaneous hand gestures in the context of emotion recognition. It can lead to a false interpretation of the emotional state because wearing something special increases the level of stress and makes the user feel discomfort. Moreover, some sensors are quite expensive.

In fact, with the evolution of open-source software libraries, many algorithms were developed based on computer vision methods using Kinect or simple camera as presented in the section 2.3.3. However, in many cases the addition of several conditions like hand is the only object with skin color in the scene or applying a primary classification, hand is the dominant object in front of the webcam, or fix a static background degrades the performance and robustness of an approach and makes use of such application more limited.

Body gestures techniques are divided into two categories [132]:

- Static gesture techniques extract features from one frame that represents the Apex frame (key frame) in a video sequence.

- Dynamic gestures techniques are based on the movement between successive frames to determine the body paths.

In our thesis we concentrate on dynamic gesture techniques that prove their efficiency in comparison to static ones [132].
We propose a hybrid approach that combines OF and HMI and obtain a motion representation of hand that will be the input of deep stacked auto-encoder (SSAE). Then, we take into consideration the relation between hand and face and we suggest a deep spatio-temporal approach based on the concatenation of temporal normalization method (EBMI) and deep learning method (SSAE) to classify emotions. In many cases when there are similarities in body gestures representations using a part-based model, confusion could happen in the body gestures interpretations, so we need to precise exactly the position of joints in body parts. In this context we propose our deep multi-stage approach based on the kinematic-based model for emotional body gesture recognition and the results are competitive over existing methods in the state of the art.
Speech is another modality for psychology recognition presented in our thesis. Several techniques exist in literature for SER and the first issue of all these techniques is the selection of the best acoustic features that could be powerful to distinguish between emotions [20]. Many researchers applied a combination of different acoustic features. According to them, extracting more features leads to a good classification accuracy. However, the increase of the vector features dimension may lead to the redundancy of information and negatively affects the classification time and accuracy. In addition, the existence of different languages, speaking styles, and accents represents a big difficulty because they directly affect the extracted features and the classification of emotions. Then, we can have more than one emotion in the same speech signal, each part represents a specific emotion so that defining the boundaries between these parts is a difficult task. Moreover, emotion can sound the same when an individual expresses different emotions. So, it is very difficult to distinguish emotions between only few individuals and the small differences are not clear with a small set of data. Therefore the emotion classification accuracy is low in comparison to other modalities. In this context, we propose our deep temporal-cepstrum representation that is effective in encoding the characteristics of speech. The results obtained prove the effectiveness of our method over existing methods in the state of the art. All the details of our proposed approaches will be presented in the next chapters.

2.6 Conclusion

We presented in this chapter various techniques in literature related to the key steps of our automatic emotion recognition system. We begin with the first modality by citing different approaches of emotional body gesture recognition using the part-based model and kinematic-based model. Then we move to the second modality and we present the related work of speech emotion recognition. In the next chapters we will talk in detail about our proposed approaches for each part already cited.

Chapter 3

Part-based Model for Emotional Body Gesture Recognition

3.1 Introduction

In this chapter, we present our proposed approach for emotional body gesture recognition using the part-based model as shown in Figure 3.1. We begin by explaining the principle of our proposed method for hand motion modeling. Then, we move to present our deep spatio-temporal approach that merges the temporal normalization method and deep learning method by taking into consideration the relationships between body parts (hands and face).

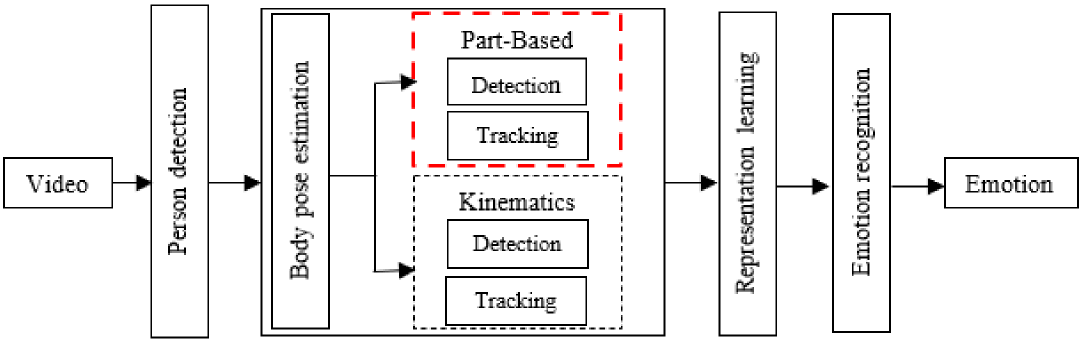

Figure 3.1: Emotional body gesture recognition system using the first scenario: After detection the person, the common step consists of estimating the body pose using the part-based model

3.2 Hand motion modeling for psychology analysis

Spontaneous and unconscious hand gestures are interesting components as they are applied to interpret the vocal content and aid observer comprehension by reinforcing the attention. Hand gesture techniques can be divided into two categories [132]: The static hand gestures techniques extract features from one frame to be classified while the dynamic ones use the movement between successive frames to determine the hand paths. In this work, we concentrate on dynamic hand gestures and we use two motion estimation techniques: Kalman Filter (KF) and Optical Flow (OF). We demonstrate that the second technique outperforms the first one so that we use it combined with History Motion Images [64] (OF-HMI) and we propose a motion representation of hand that could be useful for emotion classification.

3.2.1 Hand motion modeling using Kalman Filter method

Object tracking is interesting in the field of computer vision. The increasing need for automated video analysis and the availability of inexpensive and high-quality video cameras have generated a deal of interest in object tracking methods. Three key steps are presented in video analysis which are the detection of moving objects, tracking objects from frame to frame, and their analysis to recognize the behavior [12]. Therefore, the application of object tracking techniques is pertinent in the tasks of video indexing, traffic monitoring, motion-based recognition, human computer interaction, etc. Different methods for object tracking exist in literature [99]:

- Point tracking: KF, Multiple Hypothesis Tracking, Dense OF, Sparse OF.
- Kernel tracking: Color Histogram.
- Silhouette tracking: Contour tracking (Mean Shift),
  Shape Matching (Template Matching).

For hand motion modeling we need to determine the trajectories of hand to identify the gesture. Generally, the KF is a statistical estimation algorithm used to find the trajectory of moving object. It is essentially a set of recursive equations [146]. These equations describe the state of successive time system and with them we can predict the future state of object using the actual state. The position of the object could be corrected by considering this prediction. KF equations [77] are grouped into two equations:

$$X_k = AX_{k-1} + \alpha \tag{3.1}$$

$$Y_k = HX_k + \beta \tag{3.2}$$

Where

$X_k = [x_k, y_k, v_k, w_k]^T$ represents the vector state at instant k, (x_k, y_k) is the position of object and (v_k, w_k) is the speed vector.

$Y_k = [y_x, y_y]^T$ is the observation at instant k.

$A = \begin{bmatrix} 1 & 0 & \Delta t & 0 \\ 0 & 1 & 0 & \Delta t \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ is the matrix of state transition.

$H = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$ is the matrix of measure.

α is a random vector representing the model uncertainty.
β is a random vector representing the additive noise in the observation.
Because of its efficiency in the tracking of the object even in a complex background [146], we apply this algorithm for hand tracking as shown in Figure 3.2 by considering this hypothesis: Hand is the part that has more energy and therefore in motion. The main steps of the KF algorithm are:

1. Initialization: Initialize the position of the hand to x0 and the error tolerance weight (P0=1).
2. Prediction: Estimate the hand current state (Predict(i)) using the previous state (Actual(i)).
3. Correction: Locate the hand which is in the point predicted in the previous step. The actual position is used for making correction.

3.2.2 Overview of our proposed OF-HMI method

We propose a motion history representation of hand based on a hybrid approach that combines a technique of motion estimation of object OF and the temporal normalization method which is the HMI. We extract hand position from conversational video sequences, by exploiting the fact that hands are the most rapid part of the person's upper body eliminating the face and taking into consideration the skin color. Based on these hypotheses, hand motion modeling is a combination of face detection, skin-color segmentation binary image, face-mask image, and OF

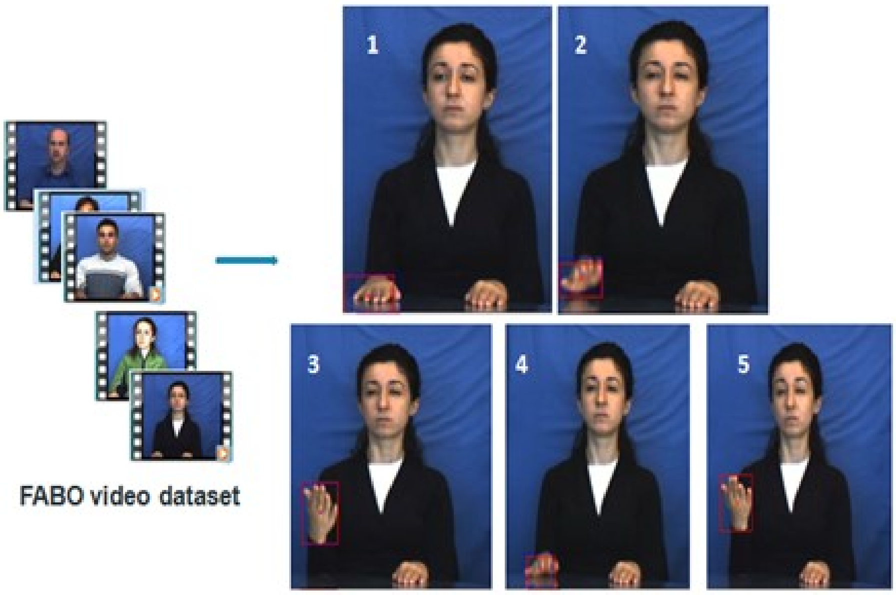

Figure 3.2: Hand motion tracking using Kalman Filter

which is a robust indicator where the hands are in conversation. An overview of our approach is presented in Figure 3.3.

3.2.3 Pre-treatment for hand detection

The first step for hand gesture recognition is the detection of two hands and the segmentation of the corresponding image regions. This segmentation is important because it separates the task-relevant data from the image background, before moving to the tracking and feature extraction for hand recognition. Plenty of effective methods exist in literature for face detection and they can give good results. In our work we choose Viola-Jones face detector [89] which is a popular method that proves its efficiency in interest area detection. This algorithm is characterized by three steps as shown in Figure 3.4.

Haar-like features A Haar-like feature consists of light and dark regions as shown in Figure 3.4. A single value was produced by taking the sum of the light regions intensities and subtracting the sum of dark regions intensities. Haar-like features allow us to extract pertinent information from the image like diagonal lines, straight lines, and edges that could be useful to identify human face. These features are computed by the integral image which accelerates the computing process. The integral image is a digital image representation. Each point (x, y) in this image contains the sum of the pixels located above and to the left of the point (x, y) of the original image.

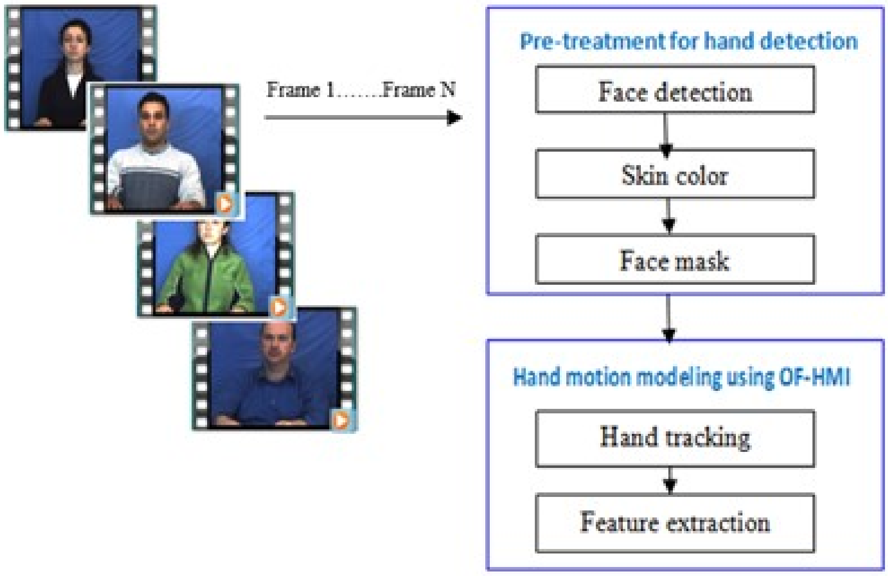

Figure 3.3: Overview of our OF-HMI architecture

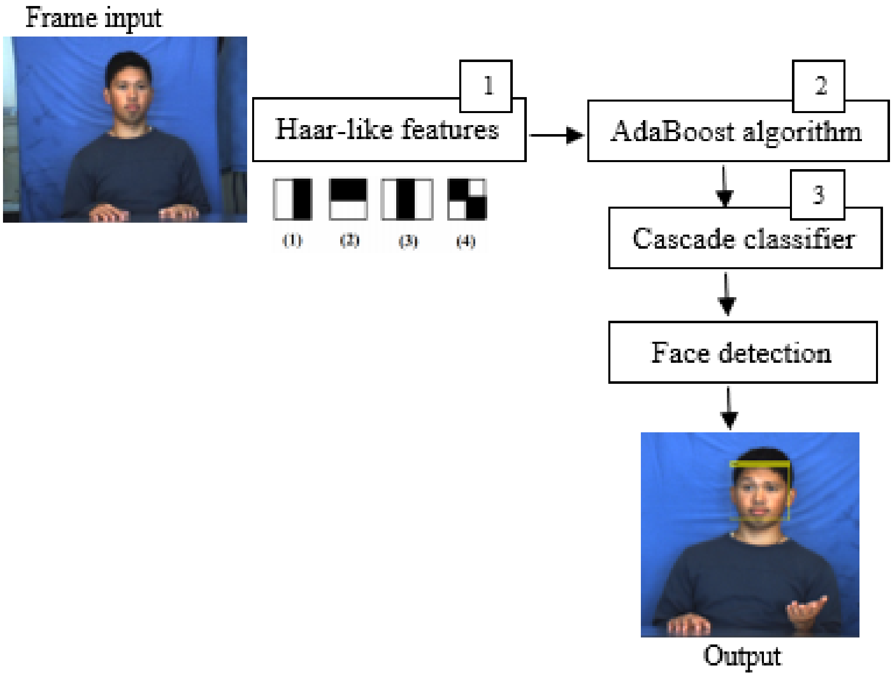

Figure 3.4: Face detection process

Adaboost algorithm Adaboost is a learning technique allowing the stimulation of the performance of a weak classifier to build a strong classifier, select, and combine the set of the weakest classifiers characterized by the best performance. The principle is to pass a set of samples to be classified through a sequence of weak classifiers. Each classifier is trained by considering the misclassified samples of the previous classifier. It uses only one characteristic at a time. Its learning phase consists of finding a threshold value of the feature making it possible to better differentiate the positive and negative objects. Indeed, each weak classifier is associated with a pseudo-Haar feature and its threshold to be able to classify an image according to the value of its characteristic. It is based on combining weak classifiers to obtain a strong one.

Cascade classifier Viola-Jones uses the technique of cascade classifier which is a multi-stage classifier that can perform the detection quickly. The cascade is composed of a series of stages and Viola-Jones learned 32 stages. Each stage contains a strong classifier produced by Adaboost. The principle of this technique is to check the classifier decision:

- Negative answer $A < 0$ (no face) if the object does not exist in the sub-window.
- Positive answer $A > 0$ (face) if the sub-window contains the object to be detected. The sub-window is forwarded onto the next stage, and it will be rejected if the answer is negative (least possible calculation). The example to classify is positive if all the stages respond positively.

Then we apply a skin color algorithm in each frame of the video sequence in YCbCr space to obtain a binary image as shown in Figure 3.5. As the hands and face are similar in skin color, we construct a face mask by setting to 0 inside the bounding box which is the face region of interest. So, with this pre-treatment we eliminate the face which is the most noticeable part that totally disrupts the tracking.

3.2.4 Hand tracking using Optical Flow

The Horn Schunck optical flow (HSOF) [121] is a technique used for motion estimation that consists of studying the position of objects in a video sequence and seeking the correlation between two successive frames to predict the change of displacement of the content and to transfer the three dimensional (3D+time) objects to a (2D+time) case. Therefore, we identify and quantify the motion of hands in interview video stream using HSOF and the result is a representation of the magnitude and direction of OF [146] at each location. This technique

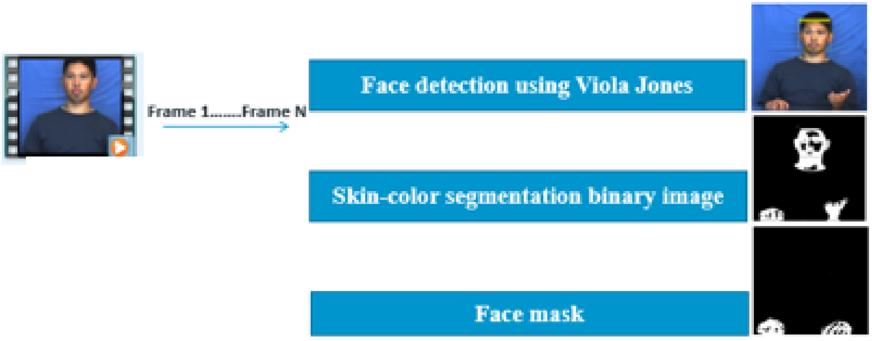
Frame 1.......Frame N
Face detection using Viola Jones
Skin-color segmentation binary image
Face mask

Figure 3.5: Pre-treatment for multiple hand detection

calculates the motion between two image frames at times t and t + dt at every position. A voxel at location (x, y, t) with intensity I (x, y, t) will be transferred by dx, dy, and dt. The flow estimation can be expressed as follows:

$$I(x, y, t) = I(x + dx, y + dy, t + dt) \tag{3.3}$$

Using Taylor series, we obtain:

$$I(x + dx, y + dy, t + dt) = I(x, y, t) + \frac{\partial I}{\partial x}dx + \frac{\partial I}{\partial y}dy + \frac{\partial I}{\partial t}dt + H.O.T \tag{3.4}$$

From these two equations and with neglecting H.O.T (Higher Order Terms) we find:

$$\frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial t}\frac{dt}{dt} = 0 \tag{3.5}$$

That results in:

$$I_X V_X + I_Y V_Y = -I_t \tag{3.6}$$

Where V_X and V_y are the x and y components of the velocity and $\frac{\partial I}{\partial x}$, $\frac{\partial I}{\partial y}$ and $\frac{\partial I}{\partial t}$ are the derivatives of the image at (x,y,t) replaced with I_X ,I_Y and I_t.
Figure 3.6 presents an example of hand tracking in the video sequence using the OF method.

3.2.5 Feature extraction

HMI is proposed initially by Davis and Bobick [94], it is a commonly used spatio-temporal representation of movement that transforms the 3D space-time information into 2D intensity image. It is a robust method for feature extraction

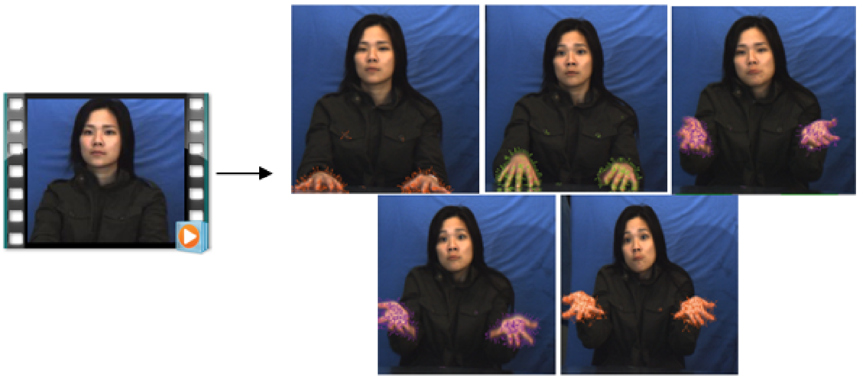

Figure 3.6: Hand tracking using Optical Flow

because it uses the depth of information, so the modeling of the hand gesture was made possible. This technique combines the gesture sequences into a single binary image using the following equation:

$$HMI(x, y) = \sum_{t=1}^{n} M_{xy}(t) \tag{3.7}$$

Where:
$HMI(x, y)$: History of binary image.
$M_{xy}(t)$: Binary image result at time t.
n: Total number of frames (video length).

3.2.6 Hand gesture representation using OF-HMI

To obtain the final representation of hand motion, the OF length represented by the function O(x,y) is accumulated for each pixel (x, y) over time. The resulting intensity value shows the historical movement at that location. The proposed OF-HMI is described in this expression:

$$S(x, y, t) = O(x, y, t) + S(x, y, t-1).m \tag{3.8}$$

With:
O(x,y,t): OF length of pixel (x, y) of frame at time t.
S(x,y,t): Image result of OF calculation in three directions for each frame at time t.
m: Update rate with 0 <m<1.
A quick moving object in a scene like waving hand requires a larger update rate but crossed arms or tapping hand on table needs a smaller rate to express the motion. The result of application of this method will be as shown in Figure 3.7.

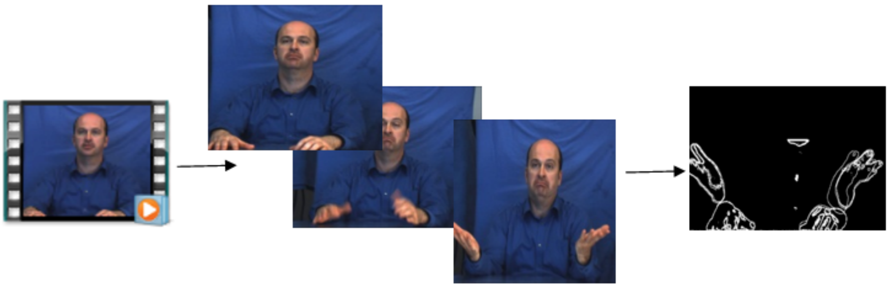

Figure 3.7: The result of application of OF-HMI in video sequence

3.3 Body gesture modeling for psychology analysis using deep spatio-temporal approach

Body postures and gestures are significant sources for transmitting emotional information when other channels are inaccessible to the observers. Taking into consideration the relationship between the body parts (hands and face) and based on the movements of the person's body parts, we classify the emotions: For example touching the chin, the neck, or the forehead reflects the uncertainty.
We propose in this work a deep spatio-temporal approach that merges the temporal normalization method with deep learning method for emotional body gesture recognition and the results prove the effectiveness of our method over existing methods in the state of the art [65].

3.3.1 Overview of our method

We propose a concept of representation to construct an informative and discriminative semantic overview for body gesture recognition system. By exploiting the fact that hands are the parts with more energy for the candidate's upper body in job interview video sequence and based on the strong relation between hands and face to express the spontaneous gestures. After key frame extraction, we move to a pre-treatment which is skin color segmentation binary image that will be after that accumulated by Energy Binary Motion Image (EBMI) to obtain a coherent local motion descriptor for each body gesture. Then, these images results will be the input of deep Stacked Auto Encoder that allows the learning of high-level information from a large number of unlabelled images patches. An overview of our architecture is presented in Figure 3.8.

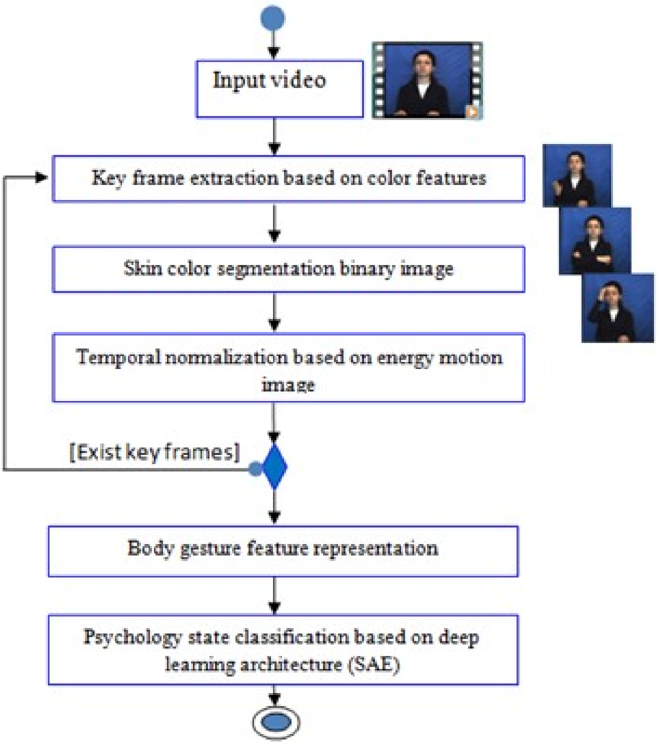
Input video
Key frame extraction based on color features
Skin color segmentation binary image
Temporal normalization based on energy motion image
[Exist key frames]
Body gesture feature representation
Psychology state classification based on deep learning architecture (SAE)

Figure 3.8: Overview of our deep spatio-temporal approach

3.3.2 Apex frame extraction

Video summarization is an important research field in computer vision for video browsing and content retrieval. The summarized video rises the efficiency in extracting the apex frames from each video. The techniques applied for extraction are numerous and depend on the type of features that are divided into two categories [18]:

- High level features like Likelihood Ratio, Pair Wise Pixel Comparison, or Motion Capture Data by Curve Saliency.
- Low level features like edge, color, or block correlation.

In a long sequence, we need just the candidate key frames which are the most informative and meaningful for the emotional interpretation. In our work we use the color histogram difference method as shown in Figure 3.9 which is robust for key frame extraction [166].
The idea behind this technique is that two frames with stable background and unchanging moving objects will have a small difference in their histograms. The color histogram difference $D(i, i+1)$ between two consecutive image frames X_i

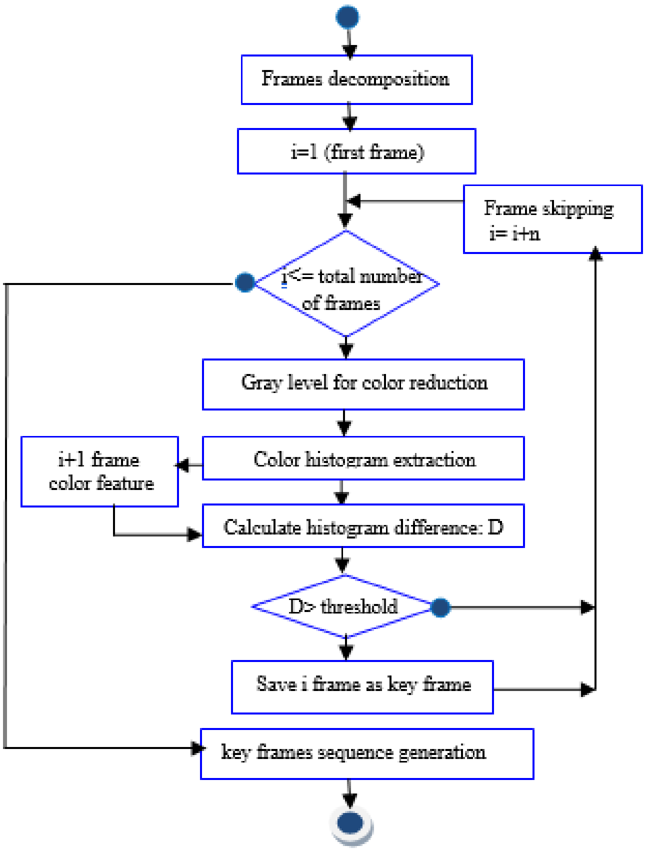
Frames decomposition
i=1 (first frame)
Frame skipping
i= i+n
i<= total number of frames
Gray level for color reduction
i+1 frame color feature
Color histogram extraction
Calculate histogram difference: D
D> threshold
Save i frame as key frame
key frames sequence generation

Figure 3.9: Pipeline of key frame extraction method

and X_{i+1} is calculated using this equation:

$$D(i, i+1) = \sum_{k=1}^{n} \frac{|H_i(k) - H_{i+1}(k)|^2}{H_{i+1}(k)} \tag{3.9}$$

With: H_i and H_{i+1} stand for the histogram of X_i and X_{i+1}. A transition occurred when the difference is bigger than a given threshold, in our case it was the sum between mean(X) and standard deviation std(X).

3.3.3 Temporal normalization for body gesture representation

By exploiting the fact that hands and head are the most energetic body parts in job interview video sequence, we develop a body gesture representation using our adaptive EBMI that can describe the spatial distribution of motion for a given view of a given gesture [94]. This method stacks the motion energy images E which are obtained through calculating the difference between consecutive key frames $(F_{j+1} - F_j)$ into a single image using the following equation and the result

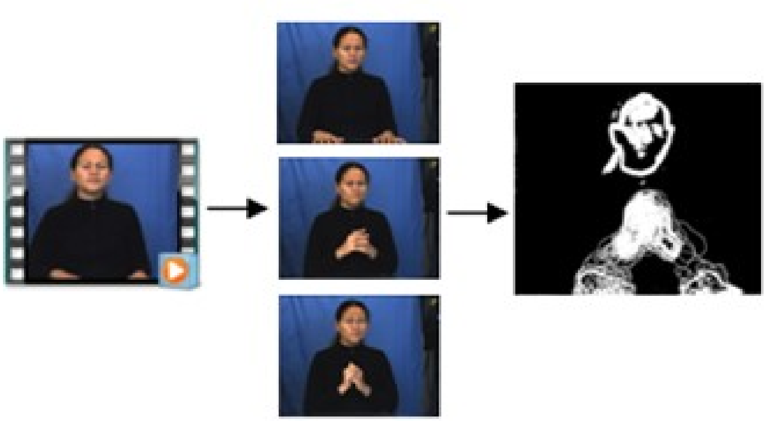

Figure 3.10: Example of application of EBMI in video sequence of anxious person

will be as shown in Figure 3.10.

$$EBMI(x,y) = \sum_{i=1}^{n} E_{xy}(i) \tag{3.10}$$

Where:
$EBMI(x,y)$: The energy of binary image result.
$E_{xy}(i)$: The energy motion image at time i.
n: Total number of key frames.

3.3.4 Emotional state classification based on deep learning architecture

Deep learning [70] is based on the use of a set of non-linear processing layers for extracting and transforming features. Thus, each layer takes as input the output of the previous one. It is characterized by a multi-level learning of details or data representations, called levels of data abstraction. Deep learning techniques prove their efficiency over traditional machine learning methods like KNN, RF, DT, and NN that need before their use the extraction of hand-crafted features (edge, texture, color, etc). However, identifying relevant features is time-consuming, computationally intensive due to high dimensions, and discriminative power is usually low. Deep learning is a powerful mechanism [11, 134] capable to learn multiple levels of data representations and to express computationally heavy models. Several architectures exist in literature. As an example, we cite the Convolutional Neural Network (CNN), Stacked Sparse Auto-Encoder (SSAE), Deep Belief Network (DBN), etc. The recent availability of powerful Graphics Processing Unit (GPU) has been an important factor in the advancement of research in deep learning. With GPU, training time for such large networks has been brought down by many orders of magnitude.

3.3.4.1 Principle of Sparse Auto-Encoder (SAE)

The success of deep learning inspired us to apply this principle of learning in the recognition of emotional body gestures, so that we propose a custom SSAE [139, 140] that benefits from the strong points of a deep network. The input layer of our Auto-Encoder (AE) consists of transforming the input image X_n into the corresponding representation $\hat{X}_n$ as shown in Figure 3.11. The hidden layer H presents the new presentation of the different functionalities of the input image patch. AE is an encoder-decoder network where the encoder represents the pixel intensities via lower dimensional attributes and the decoder reconstructs the original pixel intensities using the low dimensional features.

The weak point of AE is that it can easily memorize the training data. Therefore, regularization is necessary to overcome this problem. It gives rise to variants like sparse auto-encoders, denoising auto-encoders or contractive auto-encoders [118]. They define a simple, tractable optimization objective that can be used to monitor progress.

SAE minimizes the gap between the input images and its reconstruction in order to find the optimal parameters. For each image input x from the dataset we have:

$$TR = f_e(xW_e + m) \tag{3.11}$$

$$RE = f_d(TRW_d + n) \tag{3.12}$$

Where:
TR: transformation of the input
RE: Reconstruction input
f_e, f_d: Nonlinear activation function
W_e, W_d: Encoding and decoding weight
m, n: Decoding bias

3.3.4.2 Emotion classification using Stacked Auto-Encoder

SSAE is a neural network with multiple hidden layers of basic SAE as shown in Figure 3.12. We eliminate the part of decoding and we keep just the part of encoding in a stacked way. Our SSAE is composed of three hidden layers. First, the raw input is fed into the stacked trained Auto-Encoder to obtain the primary feature activations $h^{(1)}(x)$ for each of the input images x. Then, these primary features are used as the input to another Auto-Encoder to learn secondary features

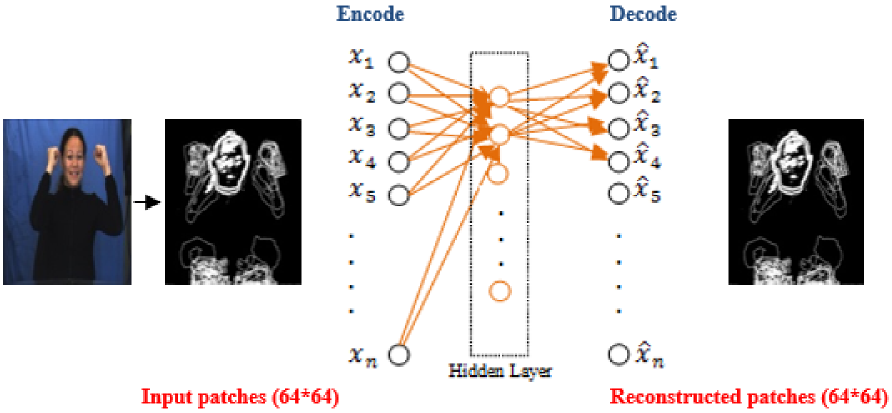

Figure 3.11: Basic of Auto-Encoder

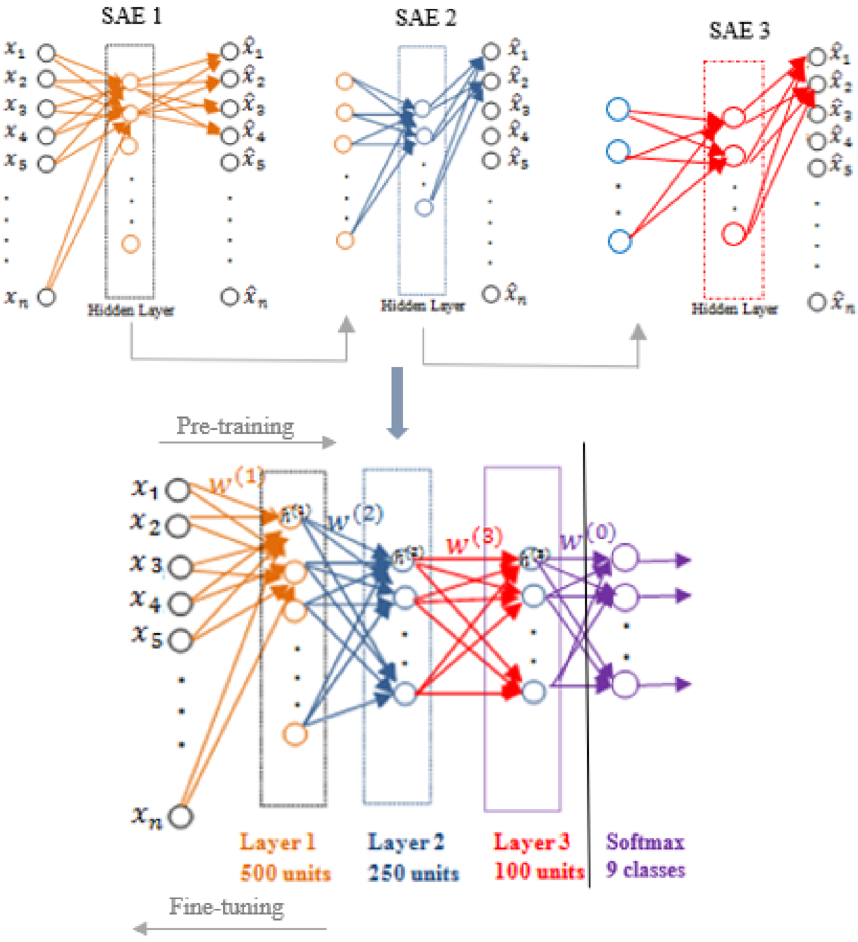

Figure 3.12: Architecture of SSAE proposed for emotion classification

$h^{(2)}(x)$ of these primary features. After that, the secondary features are entered into the third auto-encoder to get the third feature activations $h^{(3)}(x)$ for each of the secondary features $h^{(2)}(x)$. Those thirdly features are used as raw input to a Softmax classifier (SC) which is a supervised learning algorithm that has a different loss function. It can be considered as a generalization of logistic regression that we can apply for multi-class classification. Finally, to improve the performance of a stacked auto-encoder the fine-tuning with back-propagation (BP) algorithm is applied to all the hidden layers to minimize the cost function and update the weights with a labelled training set. The main advantages of BP are its flexibility in training [34] and there are no restrictions in the number of hidden layers, the number of inputs and outputs units.

3.4 Experiments and results

3.4.1 Dataset description

Directed bodily action tasks differ in appearance and timing from affective and spontaneous body gestures characterized by their freedom and liberty. So, data recording and processing for the last one is challenging. Moreover, many restrictions exist like ethical and privacy concerns and technical difficulties (illumination, consistency, etc.) that make the creation of spontaneous body gesture database more difficult [98]. Therefore, the number of datasets for emotional body gestures is still limited. Many researchers used the FABO dataset [56, 55], which is a bimodal face and body gesture database for the automatic analysis of non-verbal emotional behavior because it is well annotated and has more than 6 basic emotions. This dataset as shown in Figure 3.13 contains approximately 1900 recordings of face and body motion for 23 subjects with a length between 5 and 60 seconds (15 frames per second). In our work we are interested in the part of the dataset that captures the upper body (hands and head) to express different emotions. We split them into two thirds for training and one third for test.

FABO dataset presents ten emotional states (2 to 5 expressions for each emotion): Happiness, Sadness, Anger, Anxiety, Boredom, Disgust, Fear, Surprise, Puzzlement, and Uncertainty. We take as example: Fear, Happiness, Uncertainty, and Boredom expressions as shown in Figure 3.14, each line corresponds to an expression.

Figure 3.13: Examples from FABO dataset

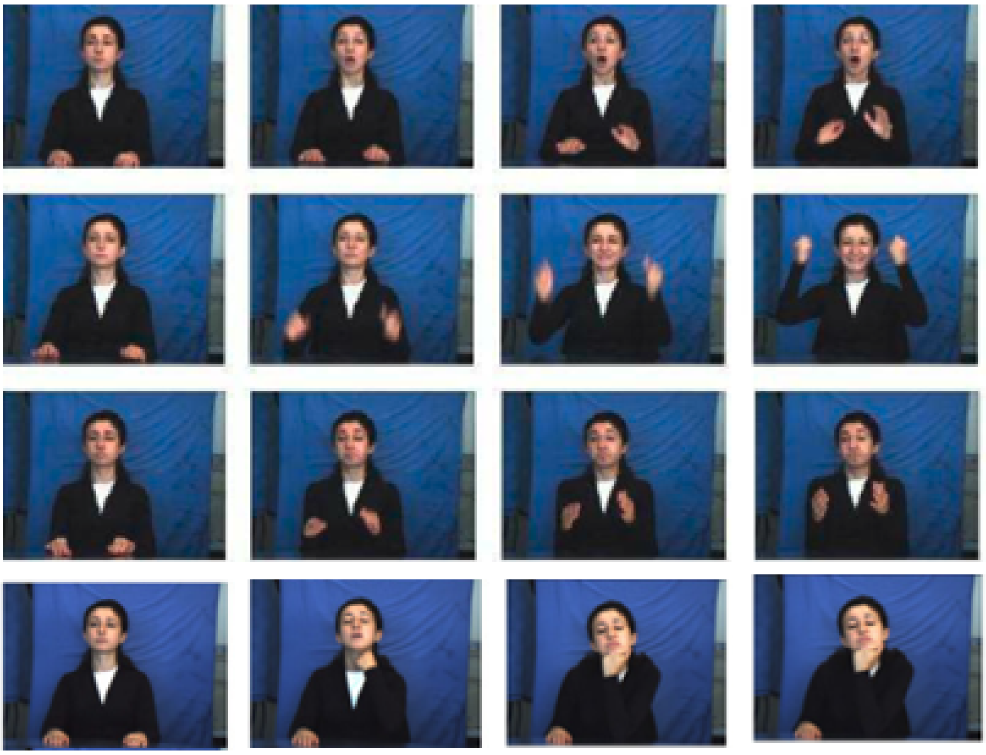

Figure 3.14: Sample images from "Fear", "Happiness", "Uncertainty", and "Boredom" expressions videos in FABO database recorded by body

3.4.2 Performance Evaluation

To evaluate the performance of our proposed methods for emotional body gesture recognition using part-based model as shown in Figure 3.1, some metrics were used which are:
Tracking rate (TR) to evaluate the tracking methods. The correctly tracked term is the number of hands which are detected and tracked correctly by the tracking algorithm and the total hand gestures term represents the total number of videos containing hand gestures:

$$TR = \frac{Correctly\ \ tracked\ \ hand}{Total\ \ hand\ \ gestures} \tag{3.13}$$

Correct classification rate (CCR) presents in our case the ratio of correctly classified hand gestures with the total number of hand gestures in the test set, misclassification rate (MCR) to check the reliability of the proposed OF-HMI for hand motion modeling. The metrics are defined as follows:

$$CCR = \frac{\sum_{j=1}^{nb} \delta(y_j, \hat{y}_j)}{nb} \times 100\% \tag{3.14}$$

$$MCR = 1 - CCR \tag{3.15}$$

Where:
$\hat{y}$: ($\hat{y}_1$,...,$\hat{y}_{nb}$) vector of predicted class labels.
δ is an indicator variable:

$$\begin{cases} \delta(y_j, \hat{y}_j) = 1 & if \quad y_j = \hat{y}_j \\ \delta(y_j, \hat{y}_j) = 0 & if \quad not \end{cases}$$

nb: total number of hand gestures in the test set.
We can also measure the performance of our method in terms of precision (Pr), recall (Re), F1 measure and Accuracy (Acc). They are defined as follows:

$$Pr = \frac{TP}{TP + FP} \times 100\% \tag{3.16}$$

$$Re = 2 \times \frac{TP}{TP + FN} \times 100\% \tag{3.17}$$

$$F1 = \frac{Re \times Pr}{Re + Pr} \times 100\% \tag{3.18}$$

Acc indicates how close the number of detections, of a specific class, is to the actual true number. We use it to compare our proposed approach with the state of the art since only this metric is reported in the corresponding publications.

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \times 100\% \tag{3.19}$$

Where TP, FP, TN, and FN represent respectively True Positive, False Positive, True Negative, and False Negative.

3.4.2.1 Hand motion modeling result

As a first experiment, we propose the KF which is an object tracking method that could be useful for hand tracking by supposing that hand is the most rapid part in a video sequence. Therefore, the center of the hand is found first and then a prediction or a correction of the position of the hand will be done in the next frame. Figure 3.15 represents an example of hand trajectory in a video sequence. We notice two curves; the red one is stable and it represents the hand position in the direction x. However for the blue curve we have two oscillations (change of position in the direction y) and it means that we have a vertical gesturing of hand repeated two times in the sequence.

Limitations: KF fails when we have more than one hand to track in a video sequence like waving with two hands or crossing arm, etc. As presented in Figure 3.16, there is a confusion in the prediction of the next position using KF when both hands have moved at the same time. We can also notice the problem of miss tracking in the following graph with interrupted trajectories in the direction x and y.

HSOF is another technique for motion estimation, which consists in studying the displacement of objects in a video sequence and seeking the correlation between two successive frames in order to predict the variation of position of the content as we present in section 3.2.4. It proves its efficiency for hand motion modeling as shown in Figure 3.17. It can control some problems of multi object tracking like appearance and disappearance of hands. Amongst, the methods used, OF algorithm is found to be more promising as it gives a better result in less computation time. We notice that OF is good for detecting moving object, similar background shapes and colors, change in velocity but it performs bad for low-resolution video as shown in Table 3.1. Using KF we have a low correct tracking rate equal to 76 % over 95 % using OF as shown in Table 3.2. Based on these experiments, we choose HSOF that gives considerable results [146] and we propose our hybrid approach that merges Optical Flow and History Motion Image as we detail in section 3.2.

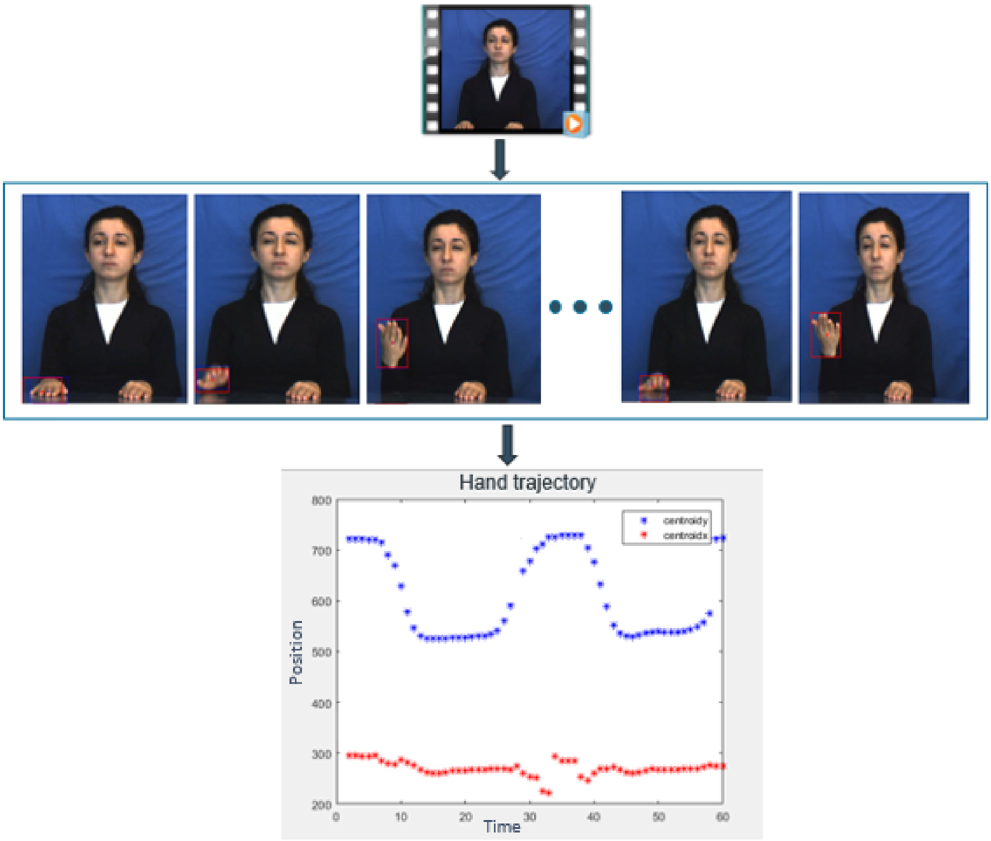

Figure 3.15: Example of hand trajectory using KF

	Kalman Filter	**Optical Flow**
Dim	Bad	Ok
Bright	Ok	Ok
Moving object	Ok	Good
Low resolution	Bad	Bad
Similar background shapes	Ok	Good
Similar background colors	Ok	Good
Change in velocity	Bad	Good

Table 3.1: Performance analysis of tracking algorithms

Algorithm	**Tracking rate**	**Time required to track hands**
Kalman Filter	76%	3-4 seconds
Optical Flow	95%	1-2 seconds

Table 3.2: Tracking rate for hand gesture

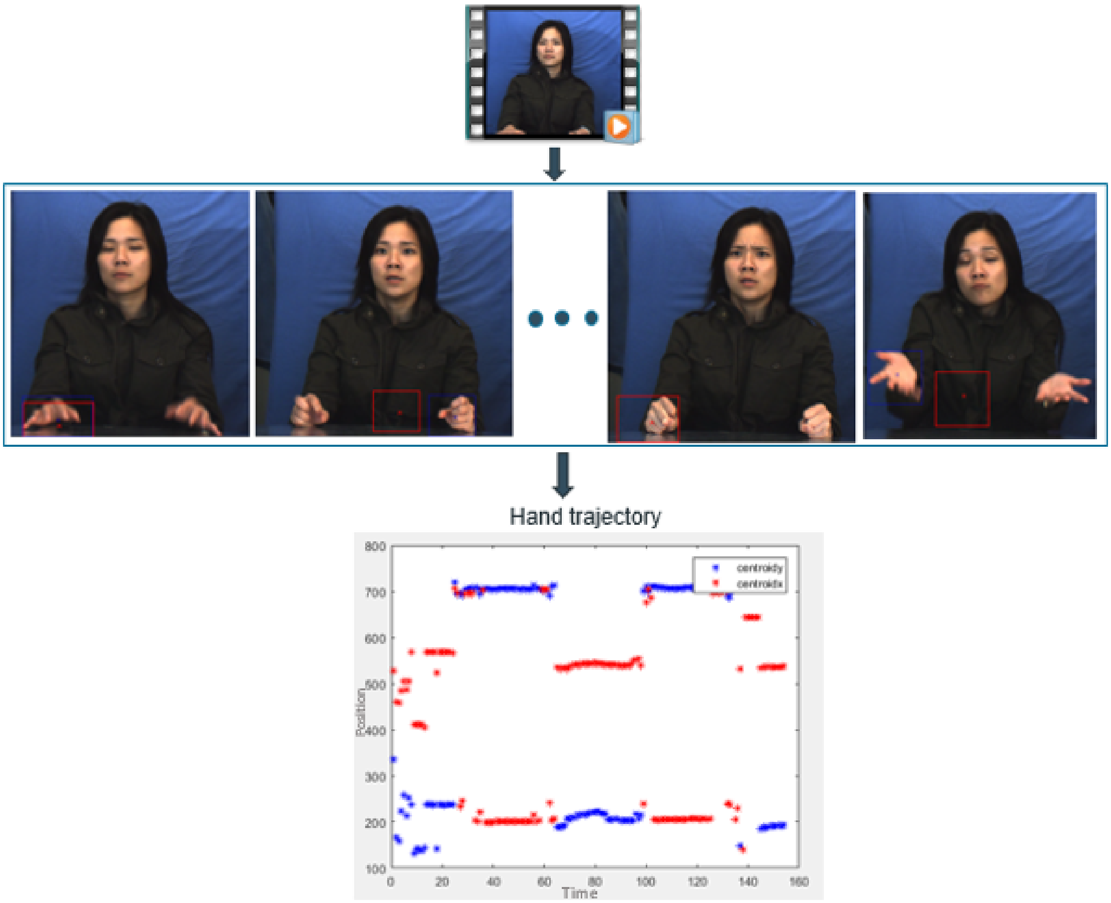

Figure 3.16: Example of miss tracking using KF

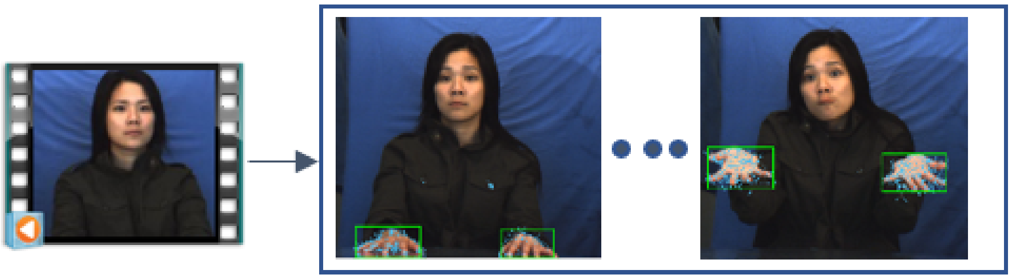

Figure 3.17: Hand tracking using HSOF

Hand gesture	**CCR**	**MCR**
Waving hand	80%	20%
Gesturing	58%	42%
Self-touch	63%	37%
Crossing arm	75%	25%
Hand on table	94%	6%

Table 3.3: Correct classification rate for hand gesture

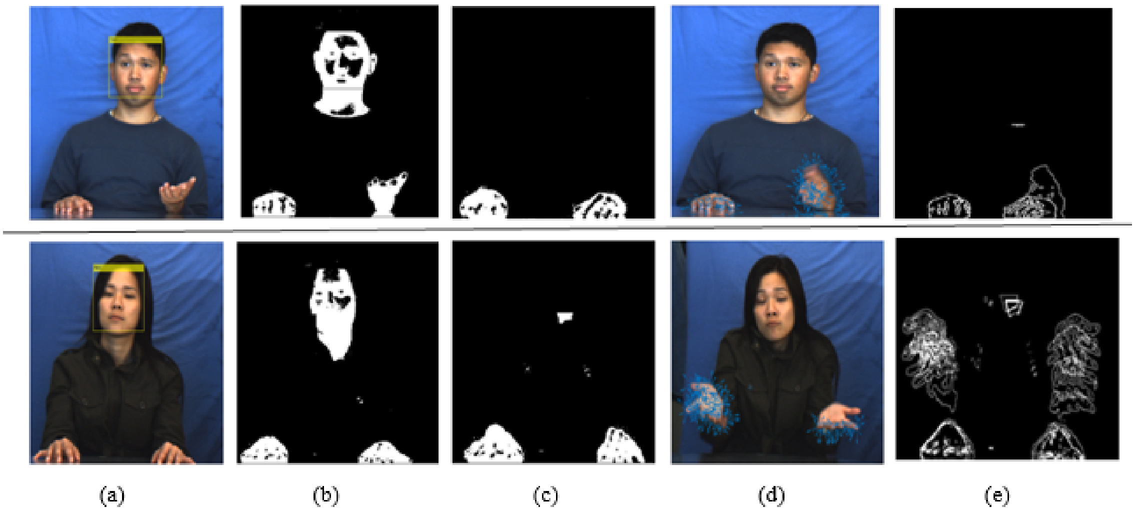

Figure 3.18: Hand motion modeling examples (gesturing and waving hand): (a) face detection, (b) skin color, (c) face mask, (d) hand tracking using HSOF, (e) motion representation using OF-HMI

Figure 3.18 represents the results obtained in the different steps of our approach implementation [64]. We begin with pre-treatments for hand detection that are; face detection, skin color to detect the regions of interest, and face mask to eliminate it because it can disrupt the hand tracking using OF. Then the outputs of OF were accumulated using HMI and we obtain a representation of the hand's local movement that will be the input of SSAE with three hidden layers: The first layer contains 500 units, the second has 250 units, and the third with 100 units. For the parameters of AEs, the Sparsity Regularization was set to 4 and the Sparsity Proportion to 0.1. Then the SC was applied to classify 5 hand gestures which are: Hand on table, crossing arm, gesturing, waving, and self-touch as shown in Table 3.3 and we obtain an overall classification rate of 75 % as shown in Figure 3.19.

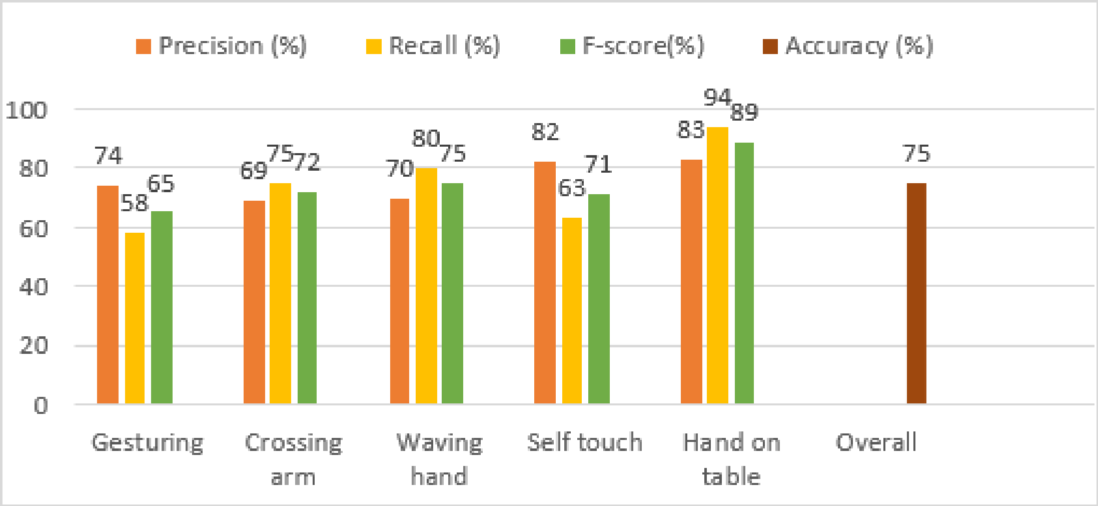

Figure 3.19: Average classification for 5 hand gestures

3.4.2.2 Body gesture modeling result

The results obtained for hand motion modeling are quite promising but not enough. According to previous studies [42, 129] the hand represents a real indicator for the psychological state interpretation but with using our proposed OF-HMI approach, a confusion could happen between gestures; for example, gesturing with one hand and self-touch have similar representations but one expresses confidence and the other expresses anxiety and stress. Therefore, we need to take into consideration the relationships between body parts (hand and face) to better differentiate between gestures. In this context, we propose our deep spatio-temporal approach as presented in the previous section.

The flowchart of our proposed method as described in section 3.3 can be summarized in Figure 3.20. We begin with key frame extraction based on color histogram difference method, and then we apply a skin color segmentation to obtain a binary image that will be accumulated using EBMI. The image result will be the input of deep SSAE with three hidden layers to classify nine emotions (happiness, sadness, anger, anxiety, boredom, disgust, fear, surprise, and uncertainty).

We compare the result obtained using our deep spatio temporal approach to other methods in the literature like Chen et al. [135]. They proposed two different methods with the same number of samples for each class of emotion from the FABO dataset: The TN method which is the MHI combined with HOG for feature extraction and SVM for classification and then the BOW with SVM.

Gunes and Piccardi [58] used the same dataset and they applied a maximum voting of apex frames for key frame extraction. They extracted features based on OF, comparison to the neutral frame, edge, and geometry. Then, they used

Method	**Number of emotions**	**Accuracy**
RF [159]	4	72%
Adaptive RF [159]	4	77.33%
MHI+HOG+SVM [135]	9	66.7%
BOW+SVM [135]	9	65.3%
SVM [58]	9	64.51%
RF [58]	9	76.87%
Our method: Key Frame+EBMI+SSAE	9	81%

Table 3.4: Comparison between deep Spatio-Temporal method and state of the art methods using FABO dataset

SVM and RF for classification .
The result obtained for emotional body gesture recognition is competitive over existing methods in the state of the art using FABO dataset with an overall accuracy of 81% and prove that our deep learning technique outperforms traditional machine learning techniques like RF and SVM as shown in Table 3.4.

3.5 Discussion

Directed bodily action tasks in controlled environments differ in appearance and timing from affective and spontaneous body gesture characterized by its freedom and liberty. Therefore, data recording and processing for the last one is challenging, and many restrictions exist that make the identification and interpretation of spontaneous body gesture more difficult in comparison to other modalities. Also, the number of datasets for emotional body gestures is still limited, so, it is very difficult to distinguish emotions between only few individuals and the differences are not clear with a small set of data. The results obtained using our proposed methods are competitive over existing methods in the state of the art but we could reach a better improvement rate using large emotional video datasets.
In many cases when there are similarities in body gestures representations using a part-based model, confusion could happen in the body gestures interpretations, so we need to precise exactly the position of joints in body parts. In this context we propose our deep multi-stage approach using a kinematic-based model for emotional body gesture recognition. This model is detailed in the next chapter.

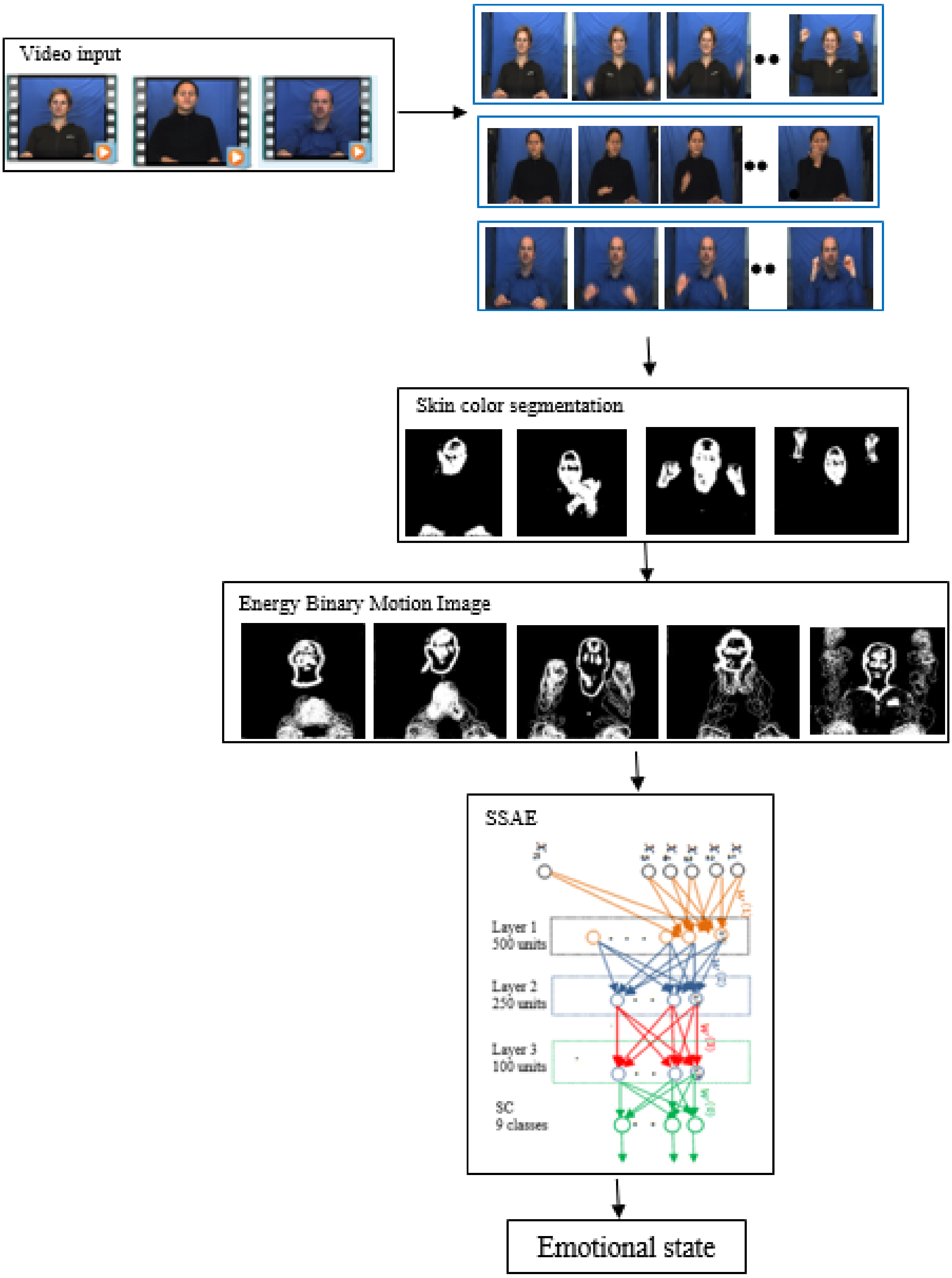

Figure 3.20: Illustration of deep spatio-temporal approach for emotional body gesture recognition

3.6 Conclusion

In this chapter, we are interested in presenting our proposed approach for emotional body gesture recognition using part-based model. We proposed a hybrid approach that incorporates two techniques which are: Motion estimation technique OF and temporal normalization method HMI to obtain a motion representation of hand that will be the input of deep stacked auto-encoder to classify hand gestures. Then, we took into consideration the relation between hand and face and we suggested a deep spatio-temporal approach based on the concatenation of temporal normalization method EBMI and deep learning method SSAE to classify emotions. In this part, we demonstrated that deep learning techniques outperform traditional machine learning techniques.

Chapter 4

Kinematic-based model for emotional body gesture recognition

4.1 Introduction

Due to the high degree of freedom, the dimension of search space, the variation of cluttered background, the body parameters and illumination; human pose estimation based on skeletal structure is still a complex task [40, 9]. The detection of body parts separately cannot be sometimes helpful for pose estimation and then for the psychology interpretation. We take the example of a binary image representation of hand and face [65] without studying the interconnection between them; a confusion could happen between two gestures like waving hand when it will be near the face and touching face. They are similar in representation, but they express different emotions: The first gesture reflects confidence and the other one reflects stress and anxiety. Kinect in this task could be helpful for the detection and then for the classification. However, it can fail when the whole body is not present in front of the camera and we have spontaneous body gestures characterized by their freedom. In this context, we present in this chapter our deep multi-stage approach. In the first task, we estimate the pose using the kinematic model as shown in Figure 4.1 to precise exactly the position of joints in the upper body. The aim of this part is to improve the 2D human pose representation through using a new structure of skeleton features based on transfer learning model rather than using the skeleton information obtained from the Kinect tool. Then, we move to the task of emotional body gesture classification.

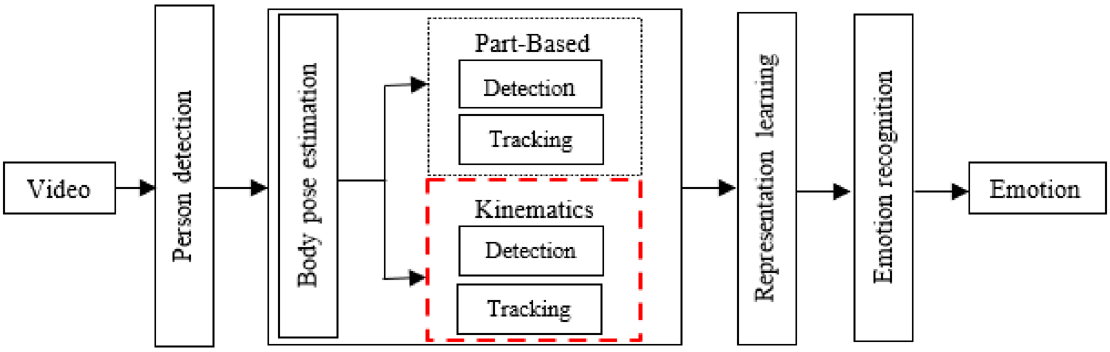

Figure 4.1: Emotional body gesture recognition system using the second scenario: Kinematic-based model

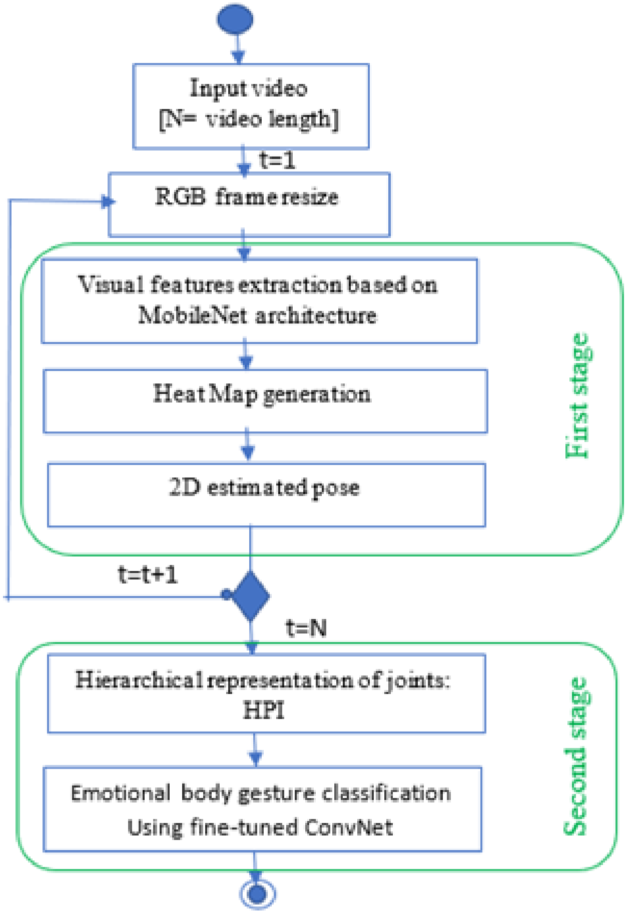
Input video
[N= video length]
t=1
RGB frame resize
Visual features extraction based on MobileNet architecture
Heat Map generation
2D estimated pose
First stage
t=t+1
t=N
Hierarchical representation of joints: HPI
Emotional body gesture classification Using fine-tuned ConvNet
Second stage

Figure 4.2: Proposed multi-stage approach

4.2 Overview of our deep multi-stage method

In this thesis, we combine two tasks of pose estimation and emotion classification for emotional body gesture recognition to propose a deep multi-stage architecture able to deal with both tasks as shown in Figure 4.2. In the first stage, our deep pose decoding method detects and tracks the candidate's skeleton in a video by fusing Depthwise Convolutional Network (DCN) which is MobileNet and detection-based method for 2D pose reconstruction. In the second stage, we propose a representation technique based on the superposition of skeletons to generate for each video sequence a single image synthesizing the different poses of the subject. We call this image: "History Pose Image" (HPI), it is used as input to CNN model based on Visual Geometry Group (VGG) architecture.

4.3 2D skeleton reconstruction

Pose estimation has been used generally for actions and activities recognition [40, 54] and proved its efficiency. Some researchers tried to refine the heat maps to improve the capacity of learning skeleton structures. Two families of pose estimation methods exist in literature:
- Detection-based methods [1] by generating the heat map, we detect the score for each corresponding joint and then the argmax function was used as post-processing to generate the coordinates of joints.
- The regression methods are based on non-linear function that directly takes the inputs to the desired outputs which are the coordinates. Based on these methods we implement our deep pose decoding method. An overview of this method is presented as follows:

- We download the weights of MobileNet model that was pre-trained on COCO keypoint dataset [10] and took as input RGB images of people annotated with keypoints.
- We prepare the input of the network by resizing the image frames.
- We make the predictions by generating the feature map.
- We calculate the confidence score by applying a sigmoid function for each joint point and then we take the high score not less than a threshold to reduce as much as possible the false detection.
- From the heat map score and based on the argmax function, we detect the corresponding coordinates of joints (the position in axis x and y).
- We draw the skeleton for each frame in the video sequence based on these coordinates.

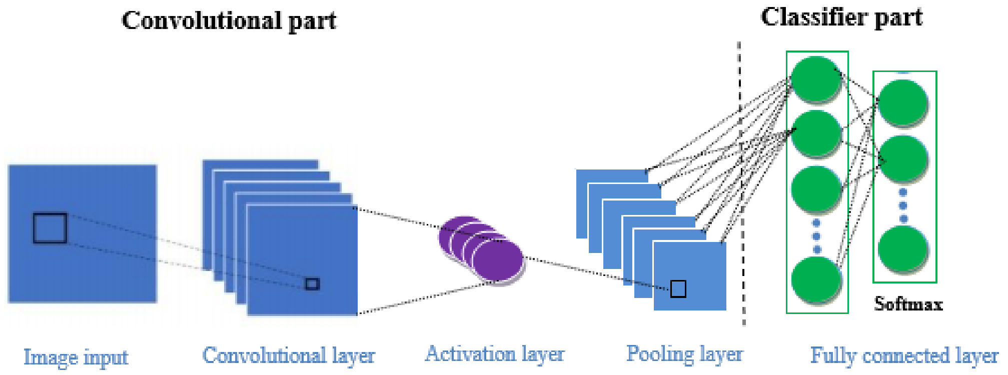

Figure 4.3: General structure of CNN

The details of our approach are explained as follows.

4.3.1 Visual features extraction based on deep learning structure

Transfer learning is a very useful method in the resolution of many tasks and especially the classification one. Instead of beginning the learning process from scratch, we fine-tune a pre-trained model and construct our deep CNN in a time-saving way. We can directly use the weights and architecture obtained of models like VGG-16 [78], ResNet [76], and MobileNet [47], which have been previously trained on large datasets like ImageNet [6] and COCO and apply the learning on our problem of pose estimation and skeleton tracking. In the last decade, some models were presented and accessible thanks to deep learning frameworks like Tensorflow, Keras, Pytorch, Caffe, etc. They are generally based on CNN [133, 26]; we talk about Convolutive pre-trained networks.

4.3.1.1 General architecture of CNN

The general architecture of the CNN is composed of four layers which are: Convolution, activation, pooling, and fully connected layer as shown in Figure 4.3. The convolution layer is the basic element of the network. It is responsible for processing the input of a receiving field. This layer contains a set of neurons that are connected to a sub-region of the preceding layer.
Then, we find the pooling (subsampling) layer between two convolution layers, it is used to reduce the spatial size of an image so the reduction of the computational cost (number of parameters and calculation) could be possible and we can avoid the phenomenon of over-learning. The most used pooling methods in the CNN are the max-pooling and average pooling. The output of max-pooling layer is given by taking the maximum value of each region of size 2×2 in the input

layer and the output of average pooling is calculated by the average value of each region of size 2 × 2 in the input layer.

The activation layer exists between the convolution layer and pooling layer, it applies activation functions to the output of convolution layer. Several mathematical activation functions exist in literature like: Rectified Linear Unit (ReLU), Leaky ReLU, Sigmoid function, etc. In our work we choose the ReLU function because it proves its efficiency [6]: It can accelerate the network construction without making a difference to the generalization of precision.

The last layer of the CNN is the fully connected layer. In fact, we can add more than one layer to improve the classification performance. When we have a supervised learning like in our case the number of neurons of the last fully connected layer is equal to the number of the desired classes. It is generally followed by a Softmax activation function that generates the probabilities distribution of classes.

In fact, the creation of the CNN from scratch is expensive in terms of hardware, expertise and the number of annotated data needed. Therefore, transfer learning could be a powerful solution by adapting publicly available pre-trained networks like the models that we use in this work: MobileNet [47] for skeleton detection and VGG [78]for emotion classification.

4.3.1.2 Skeleton features extraction based on MobileNet model

MobileNets are a new class of CNN designed recently by researchers at Google [47]. They were used firstly for mobile and embedded vision applications for the task of classification, object detection, segmentation and geo-localization, and they prove their efficiency in huge dataset like ImageNet. For example, MobileNet [47] is more accurate than GoogleNet with 2.5 times less computation. This network comprises in total 4 million (M) parameters, it is smaller and faster than others like ResNet [76] with more than 25 M parameters and AlexNet [6] with 60 M.

This model takes as input an image with the size of 224 × 224 × 3 channels, the architecture consists of a regular 3 × 3 convolution as the first layer, followed by 13 Depthwise Separable Convolution (DSC) blocks and there are no pooling layers in between. We are interested in MobileNet base network that represents the part of feature extraction. According to authors [84] the computational cost of convolution for our desired feature map will be very reduced by applying the DSC. It is a factorized convolution composed of two stages (filtering and combination stage). The Depthwise Convolution (DC) filters the inputs F and then Pointwise Convolution creates a linear combination of the output of depthwise layer to have

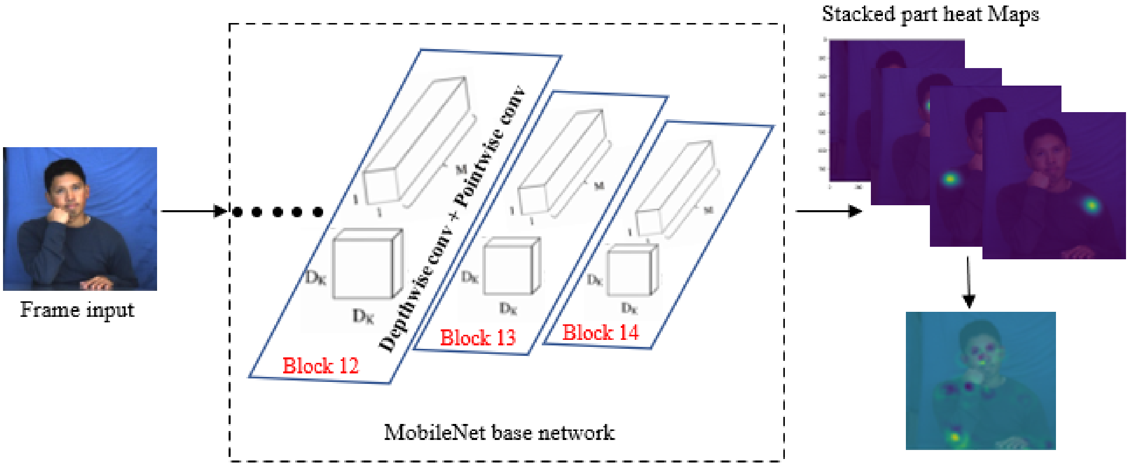

Figure 4.4: Feature extraction and heat map generation

finally the output feature map O.
The DC with one kernel for each input channel can be expressed as:

$$\hat{O}_{k,l,m} = \sum_{i,j} \hat{K}_{i,j,m} \times F_{k+i-1,l+j-1,m} \tag{4.1}$$

Where $\hat{K}$ represents the depthwise convolutional kernel with the size $D_K \times D_K \times M$ where the m filter in $\hat{K}$ is applied to the m channel in the input feature map F and produce the m channel of the output feature map $\hat{O}$.
Then the computational cost of DSC is:

$$D_F \times D_F \times M \times D_K \times D_K + 1 \times 1 \times M \times N \times D_F \times D_F \tag{4.2}$$

It represents the sum of the depthwise and 1×1 pointwise convolutions.
Where:
$D_F \times D_F$: size of input feature map F.
$D_K \times D_K$: size of depthwise convolution kernel K.
O: output feature map.
M: number of input channels or input depth.
N: number of filters or number of output channels.

After the extraction of visual features, this model generates a confidence map, which is the heat map. The part-confidence maps that are generated will be stacked to have a global heat map for each frame of the video as shown in Figure 4.4.

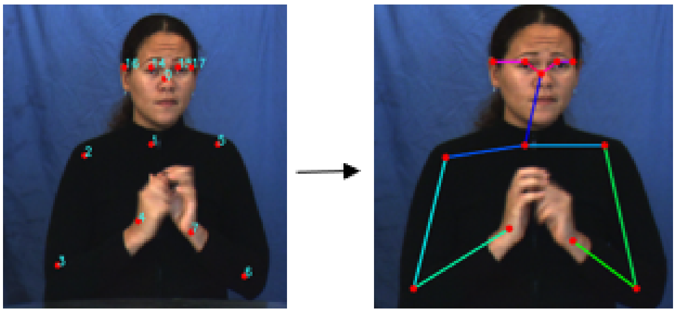

Figure 4.5: Keypoint detection and tracking

4.3.2 Detection-based approach

The main objective in the task of pose estimation is the good detection of the human skeleton in a video sequence. We compare our detected pose with the estimated pose which is the ground truth saved by the MobileNet model pre-trained on COCO dataset [10]. For each person in the image, the ground truth is composed of 18 keypoints which are nose, neck, left and right (eye, ear, shoulder, elbow, wrist, hip, knee, and ankle). In the case under analysis the candidate is in a sitting position, so we focus on upper body detection and from the 18 keypoints already labelled we detect and track the variation of the position of 12 joints as shown in Figure 4.5.

From heat map generation to 2D pose reconstruction, these are the steps of our implementation that were illustrated also in a diagram as shown in Figure 4.6. In our implementation we use some libraries offered by the programming language Python like the Open Source Computer Vision Library (OpenCV) and its functions.

Step 1. We download the model (weights and architecture) that was pre-trained on COCO keypoint dataset [165] and we load it into memory using the deep neural network (dnn) module included in OpenCV.

Step 2. We prepare the frame to be fed to the network and make predictions using the forward method for the dnn class in OpenCV. The output is a matrix with 4 dimensions: The first dimension represents the ID of image frame, the second is the index of the joint, the third is the height of the output map, and the fourth is the width.

Step 3. We calculate the confidence score by applying a sigmoid activation function for each joint point. This score is ranged between 0.0 and 1.0, and then we take the highest confidence one for each part heat map to reduce the false detection as much as possible .

Step 4. We apply a soft-argmax function to get the x and y indexes in the heat map

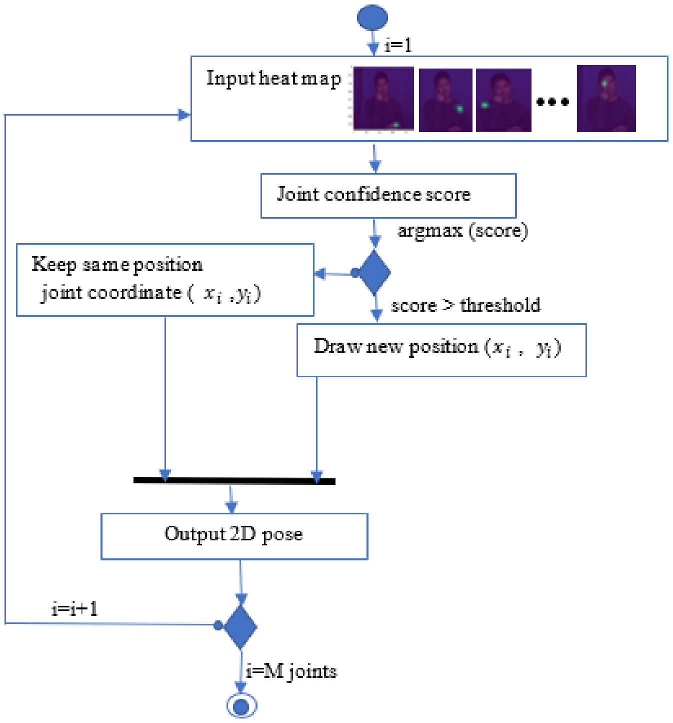

Figure 4.6: 2D pose reconstruction

with the highest score for each joint.

Step 5. When the keypoints are detected, we plot them on the image.

Step 6. We draw the skeleton by joining the pairs of keypoints as shown in Figure 4.6. If the score is more than a threshold (>=0.5) we draw the position x and y of the joint already saved in a list and if it is not, we keep the same position of joint and this treatment is done for all part heat maps of each frame.

- The treatment, from Step 2 to Step 6, were repeated for all the frames of video sequence to obtain visual and numerical hierarchical representations of joints that we detail in next section.

4.4 Human pose representation

The human representation is a challenge in the analysis of human behavior through gestures. A person in a video sequence can be described by motion, skeleton or spatial characteristics. In the last decade, some researchers proposed methods for the representation of a video in a global temporal template to convert the 3D space-time information to a 2D intensity image. The body motions of each frame are accumulated to have a single binary image. We take as examples the methods of Motion History Image (MHI) and Motion Energy image (MEI)

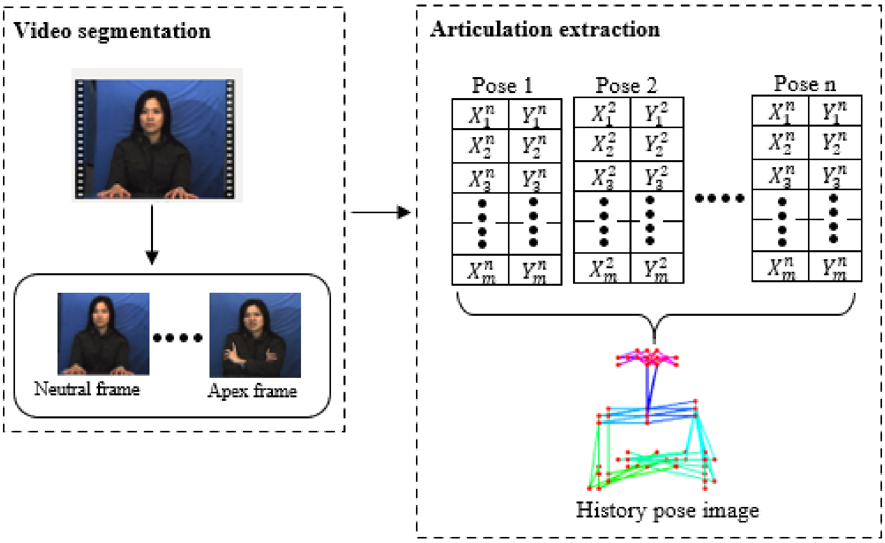

Figure 4.7: Hierarchical representation of human pose

[94, 16]: The input is a video sequence and the output is a binary image indicating the spatial location of the variations in the motion dynamics. However, the binary representation of body parts for emotional body gesture recognition could be insufficient because we want to precise the exact position of joints and their variation during the video sequence. We propose our history pose image (HPI) that stacks the skeletons detected in each frame into a single image using the equation:

$$HPI(x,y) = \sum_{t=1}^{n} M_{xy}(t) \tag{4.3}$$

Where:
$HPI(x,y)$: History pose image result.
$M_{xy}(t)$: Skeleton image at time t.
n: Number of frames.

We construct as shown in 4.7 our hierarchical representations of joints:

- The first one is numerical: We construct for each video a matrix by concatenating the different poses. This 2D poses matrix has the dimension $12 \times 2 \times number\_frames$: The height represents the number of joints and the width is the x and y position in each frame. This matrix was saved in Numpy file. It represents the emotional body gesture in a video sequence.
- The second is visual, we construct, from the points saved in the matrix, our image result called "History pose image". It will be the input of the fine-tuned VGG model to classify emotions.

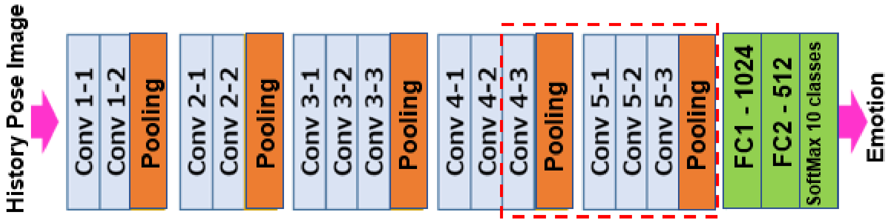

Figure 4.8: Fine-tuned VGG-16 for emotion classification

4.5 Emotion classification based on VGG architecture

In our emotion classification system, we prefer applying a transfer learning architecture since the dataset used is small and it is not efficient to build a CNN from scratch. Several pre-trained models are present in literature. The model trained on a large dataset for a source task could be a starting point to resolve our specific task. This may need the use of all parts of the model if the target dataset is big and similar to the data trained with this model. However, our dataset is small in comparison to huge ImageNet, so we need to refine and adapt the selected ConvNet, which is VGG-16 model [78, 84] that proves its efficiency in the classification task. It is a transfer learning model that has a deep architecture. The input of the model is our created 2D stacked skeleton image with the size $224 \times 224 \times 3$. We fine-tune the architecture by freezing all layers except for the last 4 layers; we retrain them because they directly affect our recognition results. Moreover, we modify the classification part by adding few dense layers: The first fully connected layer with 1024 neurons, the second with 512 neurons and the last one is a Softmax classifier to classify 10 emotions as shown in Figure 4.8. The major problem of using a model pre-trained on large dataset is the overfitting when we have limited data. Generally, the datasets for emotional body gestures are small in comparison to huge ImageNet for example. Therefore, we fine-tune and apply a Dropout regularization to improve the performance of our ConvNet as shown in Figure 4.9. When compiling the model, we optimize using an adaptive gradient descent method which is Stochastic Gradient Descent (SGD) learning optimizer [81] to accelerate our search in the direction of minimum.
Data augmentation [85] is a technique used to create variations and modified versions of images. So, the creation of new synthetic samples helps to reduce the overfitting and makes a classification system more accurate in prediction. In fact, to increase the size of our training dataset because the pose images results are limited some transformations were used:

- Rescale: Scaling our images with 1. /255 because our RGB images with

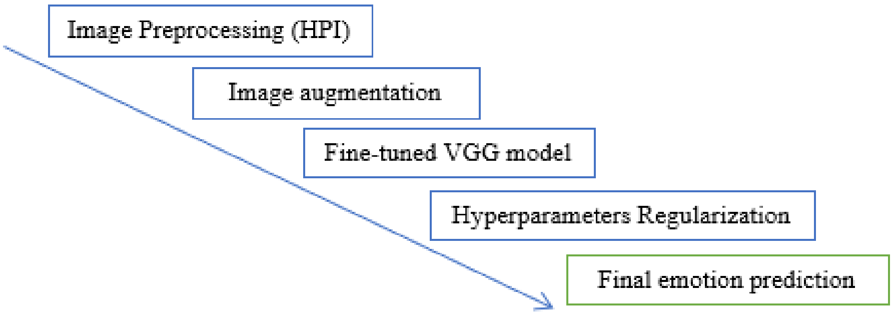

Figure 4.9: Pre-treatment for emotion classification

0-255 coefficients are high for the VGG model.

- Horizontal _flip: Takes the value True.
- Zoom _range: Zooming inside images randomly.

4.6 Experimental results

4.6.1 Dataset description

To evaluate the performance of our proposed deep multi-stage approach using the kinematic-based model, public datasets are utilized: The FABO dataset [55] is used to classify emotions (Happiness, Sadness, Anger, Anxiety, Boredom, Disgust, Fear, Surprise, Puzzlement, and Uncertainty) as presented in the previous chapter and the COCO keypoint dataset [10] is used for pose estimation (localizing the joints in body parts and drawing the skeleton). This dataset contains about 150,000 annotated people with around 1.7 million labelled keypoints.
The aim of this part is to improve the 2D human pose reconstruction through using new structure of skeleton features based on transfer learning model rather than using the skeleton information obtained from the Kinect tool. For that reason, we use the MobileNet model pre-trained on COCO keypoint dataset to localize the joints in body parts [165]. Each person in COCO has 18 keypoints as presented in Figure 4.10.
In our work, we detect the upper body of a candidate in a sitting position during an interview session. We have 12 joints: Nose (0), Neck (1), RShoulder (Right_Shoulder) (2), RElbow (3), RWrist (4), LShoulder (Left_Shoulder) (5), LElbow (6), LWrist (7), REye (14), LEye (15), REar (16), and LEar (17).

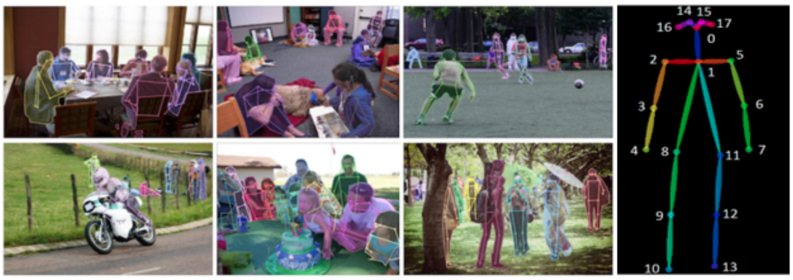

Figure 4.10: COCO dataset and skeleton with 18 keypoints

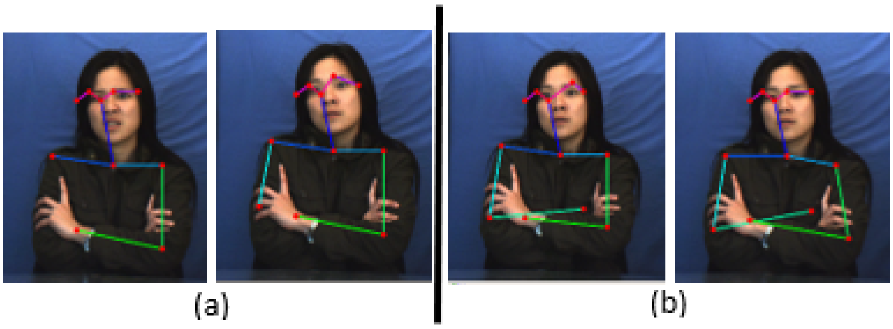

Figure 4.11: Skeleton tracking for a person with fear expression: (a) miss detection, (b) correction

4.6.2 Pose estimation result

Based on our deep pose decoding method, we detect and track the skeleton of a candidate's upper body. We calculate the corresponding coordinates (x, y) of joints and we obtain in total 24 values (12×2), taking as examples: Right_Eye_x, Right_Eye_y, Nose_x, Nose_y, etc. Sometimes some regions could be overlapped like the gesture of crossed arms: We have a confusion between two joints which are the right elbow and left wrist and vice-versa. However, this miss detection (lack of some connections between joints) of the skeleton in early frames will be auto-corrected in the next frames (see Figure 4.11): It depends on the number of keypoints detected, this number is unstable due to the variation of confidence score. If it is less than a threshold (0.5), the keypoint is missed in the frame and if it is not, the keypoint is present in the frame. Then, each skeleton detected will be stacked using our HPI method to have a single motion image which is referred to an expression. The flowchart of our implementation are presented with some examples in Figure 4.12. Each line corresponds to an emotion, the first one represents anxiety, the second is fear and the third is anger.

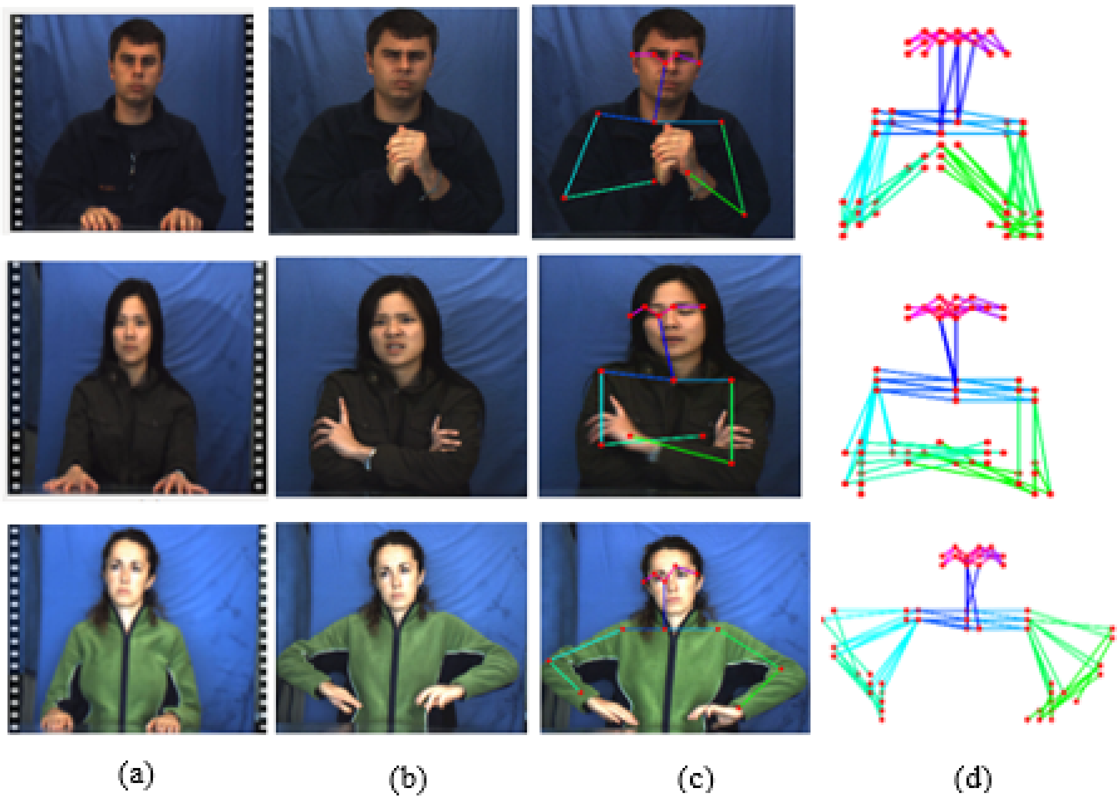

Figure 4.12: Illustration of the deep pose decoding method:
(a) video input, (b) apex frame, (c) skeleton detection, (d) history pose image

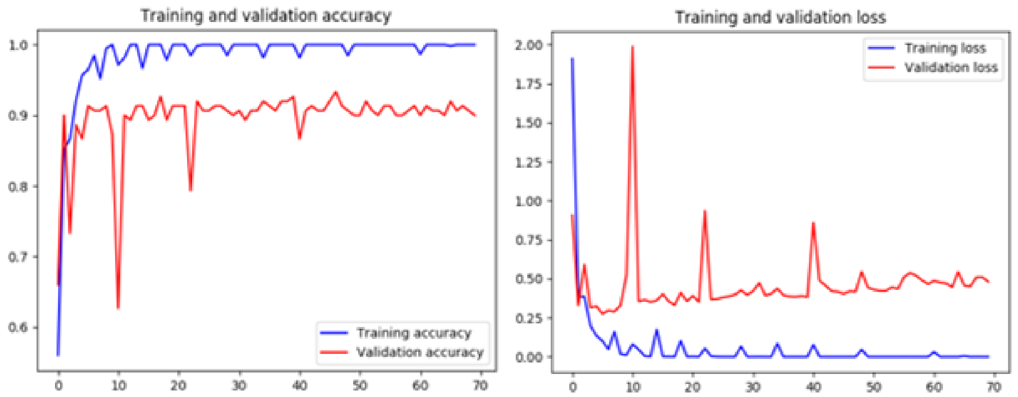

Figure 4.13: Problem of overfitting in testing accuracy graph and loss graph

4.6.3 Emotion classification results

The second stage of our proposed approach is the emotional body gesture classification. We use the fine-tuned VGG model taking as input our 2D stacked skeleton image as shown in Figure 4.12(d). After training our model with FABO dataset which is small in comparison to ImageNet, we notice a problem of overfitting and there is an acceleration in direction of the minimum. The training and validation accuracy are displayed on the graph of Figure 4.13. The loss value for the training data gradually decreases over time in a stable manner and the accuracy value for the training increases also in a stable manner and attain 89.9% for the classification of 10 emotions with 70 epochs.

However, for the validation set, the loss and accuracy suffer some unexpected oscillations: Very low picks in the accuracy graph and very high picks in the loss graph as shown in Figure 4.13. The proposed method is implemented in Keras (https://keras.io/) which is a Python-based tool for deep learning. It works on the top of Tensorflow and its purpose is the facilitation of the process of building and manipulating the neural network models. The experiments are launched on Ubuntu 16.04 computer equipped with an Intel i7-4790 CPU 3.60 GHz × 8, 16 GB RAM and NVIDIA GeForce GTX 1080 as GPU.

In fact, to ameliorate our results and overcome the problem of overfitting different functions provided by Keras could be useful:

- We add the function Dropout that drops the units from our ConvNet to the half for its co-adaptation.
- We choose SGD optimizer [81] in the compilation of the model: The learning rate (Lr) was set at $1e-4$, the decay that decreases Lr in a small factor is 0 because the Lr value is low so we don't need to modify it over each update. We also add the momentum update with a value of 0.9, it is fast when the loss function was low and vice-versa. this parameter is important for the

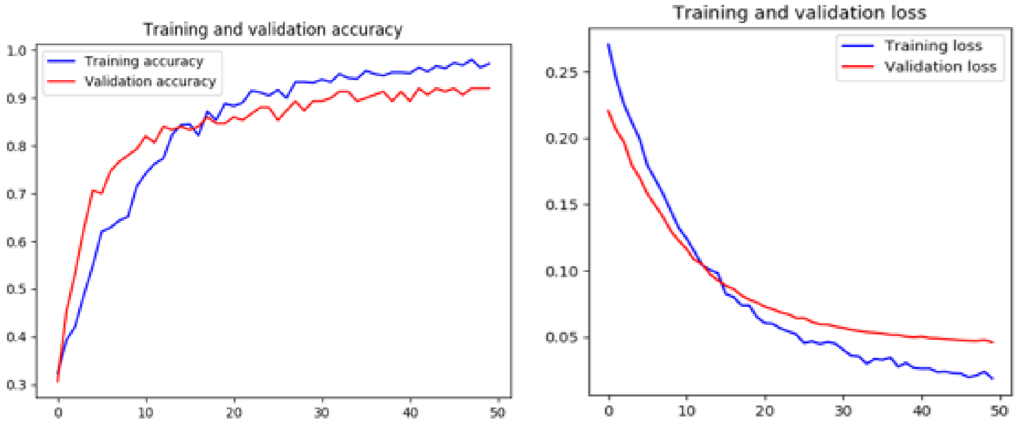

Figure 4.14: Model loss and accuracy graph for training and validation of 10 emotions

best convergence of our ConvNet because it dampens the oscillations and accumulates the gradient of the last steps to determine the next direction.

- ImageDataGenerator in Keras is a way to load and augment images in batches for image classification tasks. We create a data generator with some image augmentation: rescale, Horizontal_flip, and Zoom_range as presented in the previous section.

So, we avoid the problem of overfitting as much as possible and we obtain as a classification rate 96% in training set with a loss of 2% and 89.8% as accuracy in the testing set with a loss of 5% as shown in Figure 4.14.

Negative emotions represent real indicators for the psychological state during job interview, for example, and affect the hiring decision. We apply our proposed deep multi-stage approach to classify five negative emotions, which are anxiety, fear, disgust, uncertainty, and puzzlement. We notice an average classification accuracy of 95.37% for the five emotions as shown in Figure 4.15.

We compare the results obtained using our method to other works like Chen et al. [135] that proposed two different methods with the same number of samples for each class of emotion from the FABO dataset as shown in Figure 4.16: The TN method which is the MHI combined with HOG for feature extraction and SVM for classification and then the BOW with SVM. We notice a better classification rate for the 10 emotions like happiness with 97% and uncertainty with 93% and this is due to its clarity and distinctness in term of body gestures. However, disgust and surprise for example have less recognition rate because of the subtle nuance of expressions in gestures or maybe these emotions could be well expressed using facial cues. The comparison of our results obtained for emotional body gesture classification using a deep multi-stage approach with previous works using traditional machine learning methods and other deep learning techniques is conducted

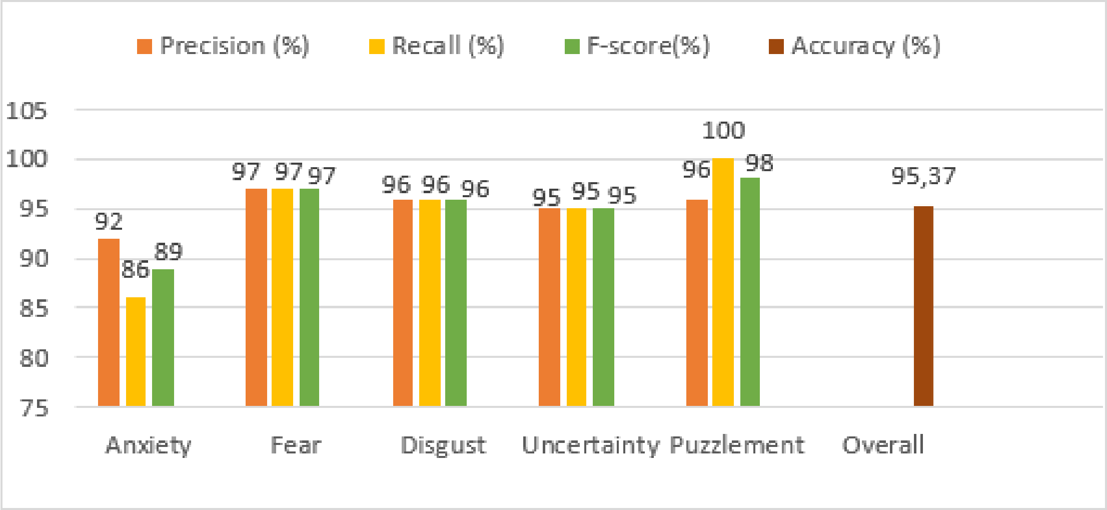

Figure 4.15: Average classification for 5 emotions

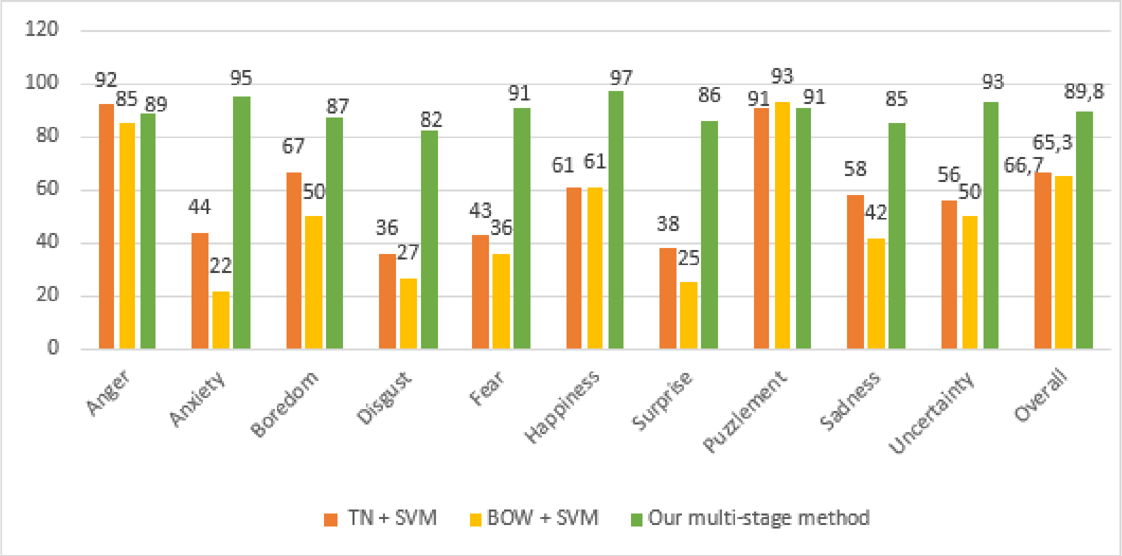

Figure 4.16: Classification rate (%) comparison of 10 emotions between our method and Chen et al. methods [135] using FABO dataset

Method	**Number of emotions**	**Accuracy%**
Chen [135]:TN(MHI+HOG)+SVM	10	66.7
Chen [135]:BOW+SVM	10	65.3
Barros [110]:CNN-gray scale	10	53.32
Barros [110]:MCCNN	10	57.84
Barros [111]:CCCNN	10	85 (2.3)
Thai Ly [147]:Key Frames+HMI	10	57.5
Thai Ly [147]:CNN+ConvLSTM	10	67.5
Thai Ly [147]:Ensemble of 2 models	10	72.5
Our	5	95.37
Our	10	89.8

Table 4.1: Comparison between our method and state of the art methods using FABO dataset

to validate the performance of the proposed approach as shown in Table 4.1.

4.7 Discussion

Body gesture conveys rich emotion but usually requires a camera pointing at the subject or wearable sensors which are not suitable for spontaneous emotion and thus for real world applications (e-health, e-learning, job interviews, etc.). Real world applications need widely available and economics camera that can be placed in the environment with minimum effort. Hence, recognizing emotion using 2D joints extracted from simple RGB video inputs is highly desirable in this context. In this part of thesis, to deal with spontaneous and unconscious body gestures for human-human communication, we propose our deep multi-stage approach based on 2D human pose estimation. The experimental results conducted in this work show the feasibility of automatic emotion recognition from sequences of body gestures. As a continuation of this work and in the next chapter we will present our proposed approach for speech emotion recognition. Speech could be an important source for transmitting emotional information when other channels are hidden (kinesics) or when there is a subtle nuance of body gesture expressions.

4.8 Conclusion

In this chapter, we presented a concept of representation to construct a semantic overview of body gestures using our multi-stage architecture. In the first stage, we estimate the pose using the proposed deep pose decoding algorithm: We extract the visual features from video sequence using a DCN model, we generate the confidence map to localize the joints in the body parts, and we draw the skeleton. In the second stage, we propose a hierarchical representation of the keypoints detected that will be the input of our ConvNet classifier.

Chapter 5

Speech emotion recognition

5.1 Introduction

In the last decade, different experiments were carried out to identify emotions from speech from different accents and languages. In this chapter, we present our deep temporal-cepstrum approach and the experiments done based on public speech datasets. The results obtained prove the efficiency of our method over existing methods in the state of the art. For SER different steps should be undertaken as shown in Figure 5.1.

5.2 Proposed deep temporal-cepstrum representation approach

For SER, the selection of the right speech descriptors is fundamental to achieve an automatic discrimination and recognition of distinct emotions. In this context, we propose a deep temporal-cepstrum representation of features that is effective in encoding the characteristics of speech (shape of the vocal tract, tone of the voice, pitch, etc.). These features are based on the concatenation of three global measures of MFCCs: static, dynamic, and acceleration part of MFCCs as shown in Figure 5.2. Then, the features are processed with our custom deep CNN for emotion classification.

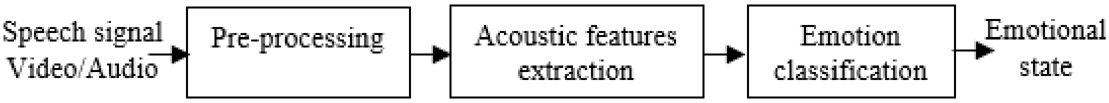

Figure 5.1: General overview of speech emotion recognition system

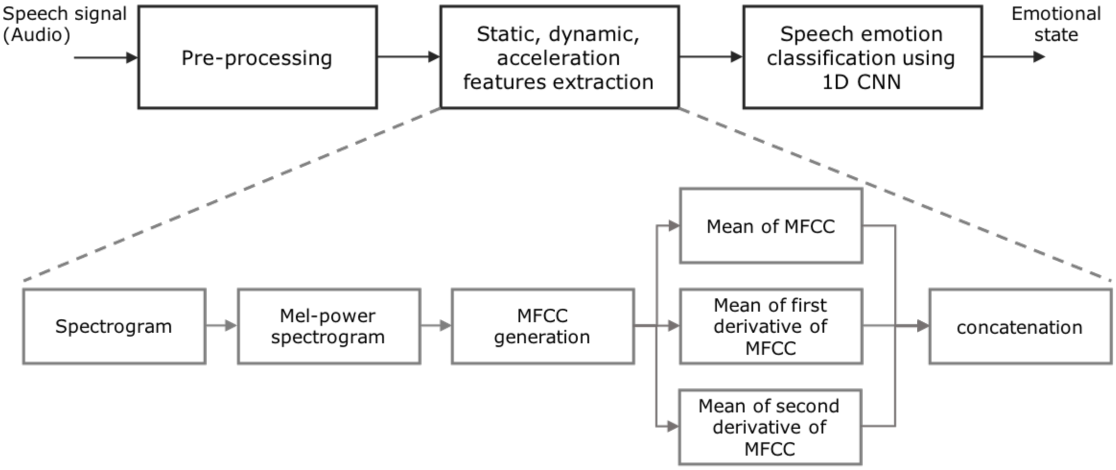
Speech signal (Audio)
Pre-processing
Static, dynamic, acceleration features extraction
Speech emotion classification using 1D CNN
Emotional state
Spectrogram
Mel-power spectrogram
MFCC generation
Mean of MFCC
Mean of first derivative of MFCC
Mean of second derivative of MFCC
concatenation

Figure 5.2: Pipeline of the proposed method:
The audio sample is initially pre-processed by removing silence and by separating it into windows of 20 ms with a step of 10 ms. For each segment of the audio sample, the Fast Fourier Transform (FFT) is applied to compute the sample spectrogram. The power of the spectrogram is mapped into the mel scale. The MFCCs are the amplitudes of the resulting spectrum. Mean of the MFCCs, of their first and second derivatives are concatenated to obtain the final feature vector. The feature vector is the input of the deep CNN. The output of the network is one of the possible emotion states, i.e. neutral, calm, happy, sad, angry, fearful, disgust and surprised.

5.2.1 Cepstral features extraction

Before feature extraction, the audio sample is pre-processed as follows:

- Silence trimming: We trim silence in each audio clip using a 30 dB threshold.
- Windowing: The input audio signal is divided into segments of 20 ms with a step of 10 ms. The Hamming window is applied to each segment in order to decrease the edge effects due to the window cutting.

After pre-processing and segmentation, MFCCs are generated since they proved to be robust in different speech recognition applications [50]. Based on comparative study [143, 104], this technique can extract the dynamic of features (linear and non-linear properties of signal), capture the important characteristics of signals because it contains both time and frequency information and it is less sensitive to noise thanks to Mel-filter bank. MFCC alone can perform well and requires less speech samples, less time consumption and reduced computationally complex compared to prosodic features [143].
MFCCs are computed as follows [62]:

- Spectrogram computation and normalization: Given an audio sample (Figure 5.3(a)), we calculate, for each segment, the magnitude of the FFT. These are then concatenated in order to obtain the spectrogram. Each frequency

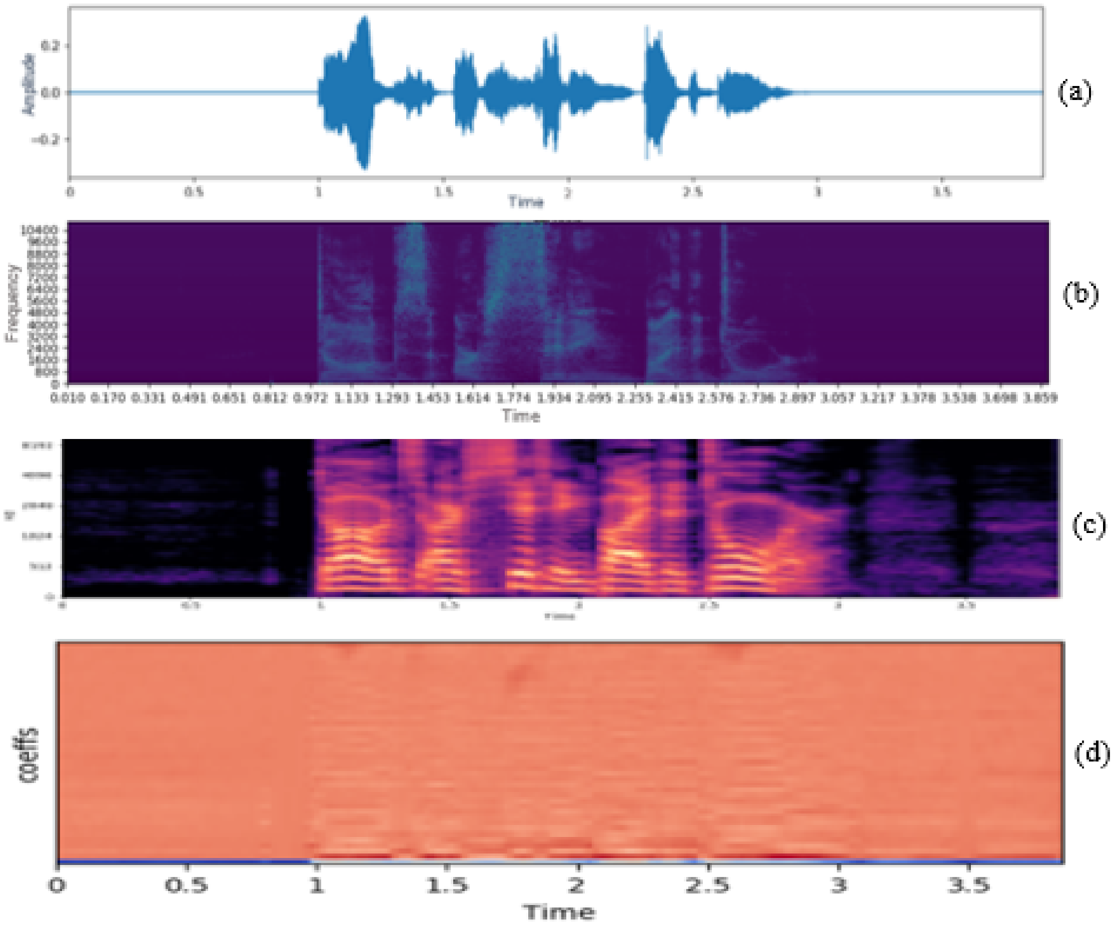

Figure 5.3: From input signal to the MFCC generation:
(a) Input signal, (b) Spectrogram normalized, (c) Mel power spectrogram, (d) 40-MFCC

bin of the spectrum is normalized by using the mean and standard deviation of the sample (see Figure 5.3(b)).

- Mel-power spectrum computation: The spectrogram power is processed with a filter bank made of triangular band pass filters. This operation maps the power of the spectrogram into the mel scale. The Mel-spectrum is the log of the output of the filters (see Figure 5.3(c)).
- MFCCs computation: The Discrete Cosine Transform (DCT) is applied to the Mel-power spectrum. The MFCCs are the first 40 coefficients of the DCT (see Figure 5.3(d)).

5.2.1.1 Static, dynamic, and acceleration features generation

The static part of our features is represented by the MFCCs while the dynamic and acceleration parts are represented by the first and the second derivatives of

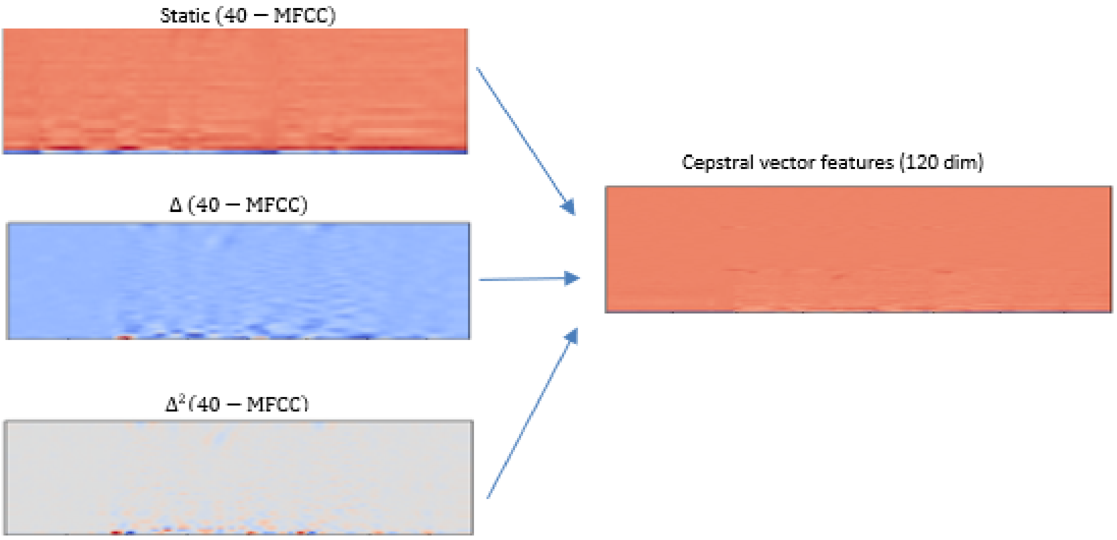

Figure 5.4: Concatenation of static, dynamic, and acceleration features

the MFCCs as shown in Figure 5.4. The temporal derivatives [24] could catch the main characteristics of the human voices that are modified by emotions, such as the shape of the vocal tract, tone of the voice, pitch, etc. In our work, the dynamic part characterizes the trajectories of the MFCCs over time and the acceleration part provides new physiology non-redundant information. For each frame t, the first derivative (Δ_t) and the second derivative (Δ_t^2) of the MFCCs are calculated as follows:

$$\Delta_t = \frac{\sum_{i=1}^{N} i(C_{t+i} - C_{t-i})}{2\sum_{i=1}^{N} i^2} \tag{5.1}$$

$$\Delta_t^2 = \frac{\sum_{i=1}^{N} i(\Delta_{t+i} - \Delta_{t-i})}{2\sum_{i=1}^{N} i^2} \tag{5.2}$$

Where N is equals to 2 and represents the number of samples considered in the calculation of the derivatives, while $\mathcal{C}_t$ is the static coefficient (coef) of the frame t of the MFCC.

5.2.1.2 Global features

Global features, such as minimum, maximum, mean, standard deviation, kurtosis, etc, have been proved to be more effective than local features in speech recognition and analysis [51]. In this paper, for each sample, we consider the mean of the MFCCs, the mean of $\Delta(MFCCs)$ and the mean of $\Delta^2(MFCCs)$. The final feature vector is the concatenation of these three components as shown in Figure 5.4, thus obtaining a 120-dimensional feature.

Layer	Input size
Conv-1D + Relu	(1, L)
Max-pooling	(L, 512)
Conv-1D + Relu	(L/4, 512)
Dropout	(L/4, 256)
Fatten	(L/4, 256)
Fully connected	(1, 7680)
Softmax	(1, 256)

Table 5.1: CNN architecture. L can be 40 or 120 on the basis of the feature vector used

5.2.2 Speech emotion recognition based on CNN classifier

Deep learning techniques have been successfully applied to many domains of affective computing [110, 147, 80] as presented in previous chapters. We design a 1-Dimensional (1-D) Convolutional Neural Network (CNN). Our static, dynamic and acceleration feature vector is the input of the CNN while the output of the network is one of the possible emotion states, i.e. neutral, calm, happy, sad, angry, fearful, disgust and surprised. The designed network is described in Table 5.1. It is composed of a convolutional layer with 512 filters having a kernel of size 5 × 1, stride 1 and padding 4. After a Relu and a pooling layer with kernel 4 × 1 we have another convolutional layer which has 256 filters. A flatten layer prepares the input for a fully connected layer (with 256 hidden neurons) which is followed by a Softmax, it is a final layer that generates the class probabilities for each speech cepstrums. The output of the network is a vector (1; K) where K is the number of emotions to be predicted. For RAVDESS dataset: K equal to 8, and for EMODB: it is equal to 7. The categorical cross-entropy is applied as loss function and Adam optimizer is used with a learning rate of $1e-3$ to minimize the loss function over mini batches of 64 speech segments of the training data.

5.3 Experimental results

5.3.1 Datasets description

To evaluate the performance of our proposed method we employ two public datasets and we split each of them into two thirds for training and one third for test.

Figure 5.5: Examples from RAVDESS dataset

5.3.1.1 Ryerson Audio-Visual Emotional Speech and Song (RAVDESS) dataset

It is composed of 7356 recordings in English [88] with an average duration of 3 seconds. Audio clips with 8 emotions with normal and strong intensity are recorded by 24 professional actors (12 males, 12 females). The emotions are neutral, calm, happy, sad, angry, fearful, disgust and surprised. All actors produced 104 distinct vocalizations that have been extracted to create three separate modality conditions: audio-only (48 kHz .wav), audio-video (48 kHz, .mp4), and video-only (no sound). In our experiments we use the 4784 recordings from the two modalities (audio-only and audio video).

5.3.1.2 The Berlin Database of Emotional Speech (EMODB)

It is a public German dataset [41] composed of 535 recording files spoken by 10 actors (5 male and 5 female). It is popular in the domain of emotion recognition due to the good quality of its recording. The average duration of the audio files is 3 seconds with sampling rate of 16 kHz. The speakers produced 7 emotions: neutral, happy, sad, angry, fearful, disgust and bored speech utterances.

5.3.2 Performance evaluation

For SER, we propose deep temporal-cepstrum representation of features, which is static, dynamic, and acceleration cepstral features combined with CNN as we presented in the previous section. We measure the performance of our method in terms of Pr, Re, F1 measure, and Acc. The comparison with the state of the art is

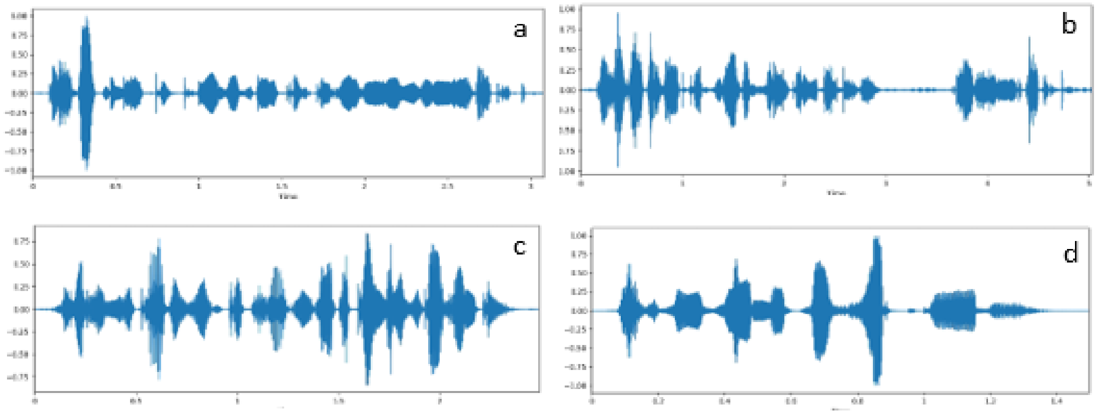

Figure 5.6: Examples from EMODB dataset:
(a) Neutral signal, (b) Sad, (c) Fearful, (d) Angry

made in terms of accuracy since only this metric is reported in the corresponding publications. The terms are defined as follows:

$$Pr = \frac{TP}{TP + FP} \times 100\% \tag{5.3}$$

$$Re = \frac{TP}{TP + FN} \times 100\% \tag{5.4}$$

$$F1 = 2 \times \frac{Re \times Pr}{Re + Pr} \tag{5.5}$$

$$Acc = \frac{TP + TN}{TP + FN + FP + TN} \times 100\% \tag{5.6}$$

To show more details about the performance of our method, we use the confusion matrix, which is a tool for measuring the quality of a classification system. The column represents the occurrences of an estimated class and the row represents the occurrences of an actual class.

5.3.2.1 Cepstral features extraction result

From each speech signal from the RAVDESS and EMODB datasets, a vector of temporal derivatives cepstral features was generated as shown in Figure 5.7(c). It is based on the concatenation of three global measures of MFCCs as shown in Figure 5.4.

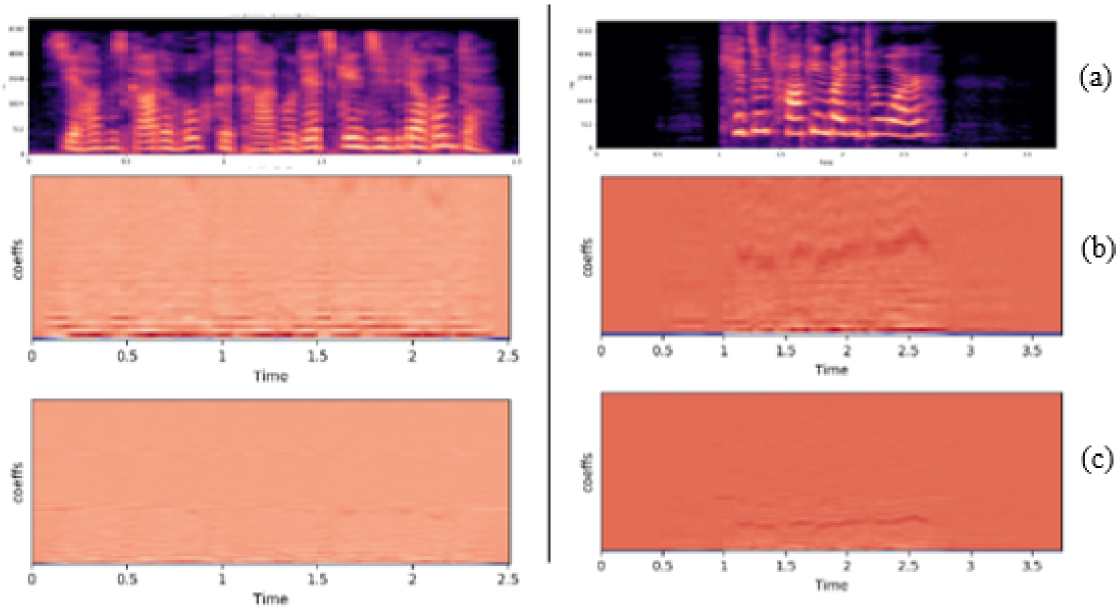

Figure 5.7: Cepstral features extraction of the speech of fearful person: On the right using RAVDESS, on the left using EMODB: (a) Mel-spectrogram, (b) 40-MFCC, (c) 120-dimentional cepstral features

5.3.2.2 Emotion recognition result

For the sake of completeness, we have also processed our proposed features using RF classifier [109]. It is an estimator that fits a number of decision tree classifiers on various sub-samples of the dataset. We get prediction from each tree and we select the best solution by means of voting. The use of averaging improves the accuracy and control overfitting. The input of RF is the $training\ sets \times nfeatures$(40 or 120). The number of trees in forest is fixed to 60 of depth 10.

Tables 5.2 and 5.4 show the results achieved using the simplified version of our features (mean of the 40 MFCCs) combined with the RF and 1-D CNN classifiers on RAVDESS and EMODB datasets, respectively. It can be noticed that the use of our 1-D CNN increases the overall accuracy of about 10% on both datasets with respect to the use of RF.

Tables 5.3 and 5.5 show the results achieved using the 120-dimensional feature vector composed of the mean of $40MFCC + 40\Delta(MFCC) + 40\Delta^2(MFCC)$ combined with RF and 1-D CNN on RAVDESS and EMODB datasets respectively. In addition, in this case, the 1-D CNN increases performance with respect RF:

- From an overall accuracy of 80.24% to 93.41% for RAVDESS dataset as presented in Table 5.3, Figure 5.8, and seen in the confusion matrix (Figure 5.9) that summarizes the results of classification.
- From 76.50 % to 86.48% for EMODB dataset as shown in Table 5.5, Figure

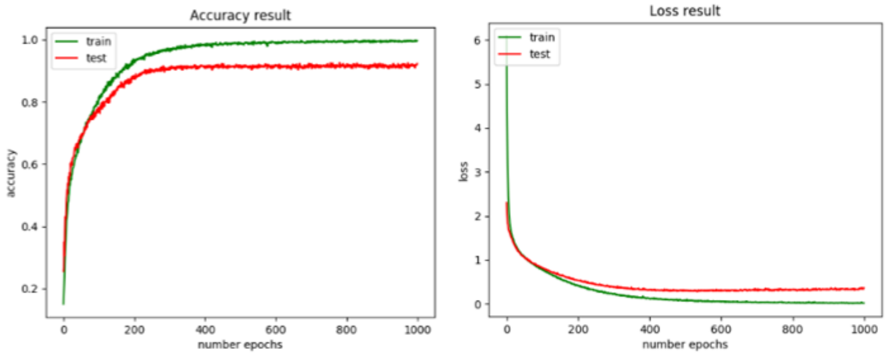

Figure 5.8: Learning curves obtained for classification of 8 emotions using RAVDESS dataset

	Metric	Neutral	Calm	Happy	Sad	Angry	Fearful	Disgust	Surprised
RF	Precision	0.83	0.86	0.78	0.77	0.82	0.76	0.73	0.86
	F-score	0.89	0.91	0.82	0.79	0.77	0.80	0.67	0.71
	Recall	0.86	0.88	0.80	0.78	0.79	0.78	0.70	0.78
	Accuracy				80.24 %				
CNN	Precision	0.92	0.90	0.89	0.85	0.95	0.95	0.96	0.95
	F-score	0.92	0.97	0.92	0.91	0.91	0.93	0.83	0.87
	Recall	0.92	0.93	0.90	0.88	0.93	0.94	0.89	0.91
	Accuracy				91.38 %				

Table 5.2: Comparison between RF and CNN using Mean MFCC (40) on the RAVDESS dataset

5.10, and the confusion matrix (Figure 5.11).

Most importantly, we mainly wanted to investigate the effectiveness of the designed features: The use of our temporal-cepstral (120-dimensional) feature vector increases the accuracy of about 6% and 2%, in the case of RF and 1-D CNN respectively, with respect to the use of the 40-dimensional feature vector. This proves that the use of dynamic and acceleration features increases the performance with respect to the use of only static features.

Finally, the comparison of our methods with the state of the art is reported in Tables 5.6 and 5.7. In these tables, the methods proposed by researchers are detailed in the chapter 2. It could be noticed that for RAVDESS dataset the improvement in accuracy with respect to the best method in the state of the art is about 15%. Concerning the EMODB dataset, the improvement in accuracy with respect to the best method in the state of the art is about 5%.
Our proposed method was implemented in Python which is a high-level programming language providing several modules and high-quality machine learning libraries. We take as example Librosa [105] that we used in this part; it is a

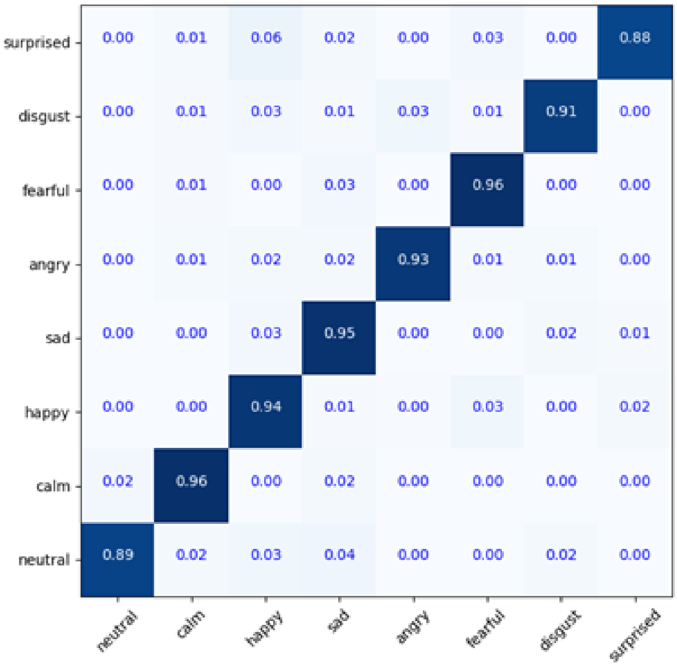

Figure 5.9: Confusion Matrix of 8 emotions using RAVDESS dataset

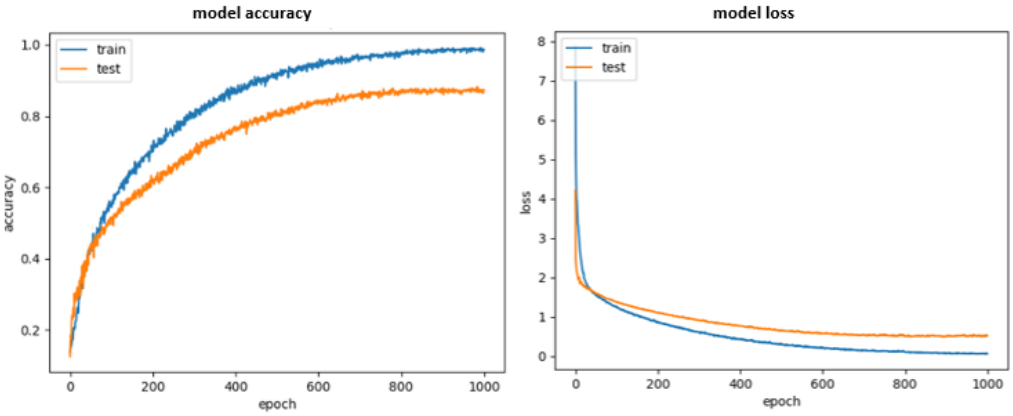

Figure 5.10: Learning curves obtained for classification of 7 emotions using EMODB dataset

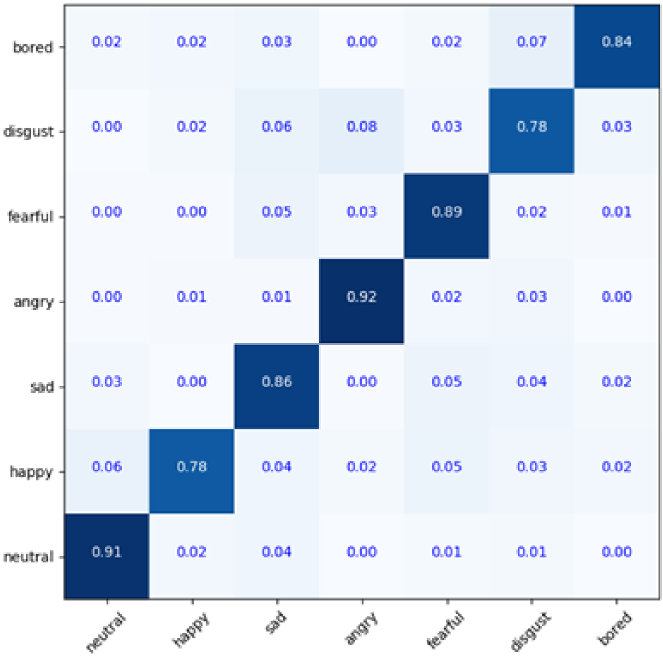

Figure 5.11: Confusion Matrix of 7 emotions using EMODB dataset

	Metric	Neutral	Calm	Happy	Sad	Angry	Fearful	Disgust	Surprised
RF	Precision	0.82	0.89	0.93	0.84	0.89	0.83	0.77	0.88
	F-score	0.91	0.93	0.89	0.86	0.89	0.88	0.69	0.75
	Recall	0.86	0.91	0.91	0.85	0.89	0.85	0.72	0.81
	Accuracy					86.40 %			
CNN	Precision	0.95	0.96	0.90	0.90	0.98	0.93	0.92	0.95
	F-score	0.89	0.96	0.94	0.95	0.93	0.96	0.91	0.88
	Recall	0.92	0.96	0.92	0.92	0.95	0.95	0.92	0.91
	Accuracy					93.41 %			

Table 5.3: Comparison between RF and CNN using our temporal derivative cepstral features on the RAVDESS dataset

package of audio signal processing that offers implementations of a variety of functions for signal decomposition, acoustic features extraction, effects, and temporal segmentation etc.

	Metric	Neutral	Happy	Sad	Angry	Fearful	Disgust	Bored
RF	Precision	0.74	0.80	0.80	0.83	0.84	0.61	0.62
	F-score	0.91	0.74	0.80	0.82	0.79	0.50	0.59
	Recall	0.82	0.77	0.80	0.83	0.81	0.55	0.61
	Accuracy	76.50 %						
CNN	Precision	0.82	0.76	0.96	0.87	0.95	0.86	0.70
	F-score	0.85	0.97	0.74	0.71	0.91	0.86	0.93
	Recall	0.84	0.86	0.83	0.78	0.93	0.86	0.80
	Accuracy	84.75 %						

Table 5.4: Comparison between RF and CNN using Mean MFCC (40) on the EMODB dataset

	Metric	Neutral	Happy	Sad	Angry	Fearful	Disgust	Bored
RF	Precision	0.78	0.87	0.82	0.90	0.87	0.67	0.76
	F-score	0.91	0.76	0.84	0.87	0.86	0.65	0.72
	Recall	0.84	0.81	0.83	0.89	0.87	0.66	0.74
	Accuracy	82.37 %						
CNN	Precision	0.91	0.93	0.82	0.92	0.85	0.70	0.87
	F-score	0.91	0.78	0.86	0.92	0.89	0.78	0.84
	Recall	0.91	0.84	0.84	0.92	0.87	0.74	0.85
	Accuracy	86.48 %						

Table 5.5: Comparison between RF and CNN using our temporal derivative cepstral features on the EMODB dataset

Method	**Features**	**Classifier**	**Accuracy (%)**
Shegocar et al.[117]	278 pros. features	Q-SVM	60.10
Popova et al. [120]	Spectrogram	VGG-16	71.00
Tanmoy et al. [155]	DWT	SVM	73.67
Tanmoy et al. [155]	DWT	GNB	77.71
Tanmoy et al. [155]	DWT	KNN	69.41
Parry et al.[119]	Spectrogram	CNN-LSTM	65.67
Zamil et al. [13]	13-dimensional MFCC	LMT	70.00
Our	40-dim MFCC	RF	80.24
Our	40-dim MFCC	CNN	91.38
Our	Mean $MFCC + \Delta + \Delta^2$	RF	86.40
Our	Mean $MFCC + \Delta + \Delta^2$	CNN	**93.41**

Table 5.6: Comparison between our method and state-of-the-art methods using RAVDESS dataset

Method	**Features**	**Classifier**	**Accuracy (%)**
Lampropoulos et al. [86]	MPEG-7 descriptors	RBF-SVM	77.88
Tanmoy et al. [155]	DWT	SVM	73.74
Tanmoy et al. [155]	DWT	GNB	80.88
Tanmoy et al. [155]	DWT	KNN	72.75
Zamil et al. [13]	13-dimentional MFCC	LMT	64.51
Parry et al. [119]	Spectrogram	LSTM	69.72
Our	40-dim MFCC	RF	76.50
Our	40-dim MFCC	CNN	84.75
Our	Mean $MFCC + \Delta + \Delta^2$	RF	82.37
Our	Mean $MFCC + \Delta + \Delta^2$	CNN	**86.48**

Table 5.7: Comparison between our method and state-of-the-art methods using EMODB dataset

5.4 Discussion

Several techniques exist in literature for SER and the first issue of all these techniques is the selection of the best acoustic features that could be powerful to distinguish between emotions [20]. In addition, the existence of different languages, speaking styles, and accents represents a big difficulty because they directly affect the extracted features and the classification of emotions. Then, we can have more than one emotion in the same speech signal, each part represents a specific emotion so that defining the boundaries between these parts is a difficult task. Moreover, emotion can sound the same when an individual expresses different emotions. So, it is very difficult to distinguish emotions between only few individuals and the differences are not clear with a small set of data. Therefore the emotion classification accuracy is low in comparison to other modalities. Despite of the small set of data, the results obtained using our deep temporal-cepstrum representation method are competitive over existing methods in the state of the art, improvement rate close to 15%, being able to better model those characteristics of the human speech which are influenced by emotions.

5.5 Conclusion

This chapter has shown the method to construct a deep short-term power spectrum representation of speech. It is based on the combination of static, dynamic and acceleration spectral features with custom CNN that combines these cepstral features coefficients with their different weights to extract the significant information for recognizing emotions such as happiness, sadness, surprise, etc. In the next chapter, we conclude with general conclusions and possible perspectives.

Chapter 6

Conclusions

At the end of this manuscript, it is now appropriate to draw up a balance sheet of our thesis research activities by shedding light on the two modalities that we have used for the task of emotion recognition and the different approaches that we have proposed.

Affective Computing is an interdisciplinary field spanning computer science and psychology. The goal is to give the machines the ability to express, recognize, and regulate emotions. Recognizing emotional information needs the extraction of significant patterns from the gathered data. This is done using machine learning techniques that process different modalities. In this dissertation, we looked in detail at the role of two modalities, visual and auditory expressions, for communicating emotions. We proposed models with the aim to replace the manual coding performed by an observer with an automatic psychology recognition system based on the analysis of non-verbal behaviors by exploiting the strong points of deep learning and transfer learning techniques.

In the first chapter a general overview of the thesis was presented. Motivations and open issues of this work were discussed. In the second chapter we talked about the review study of different works related to the key steps of our proposed emotion recognition system which are: Part-based model for emotional body gesture recognition, kinematic-based model for emotional body gesture recognition, and speech emotion recognition.

Then, in the rest of the report we detailed each step in a separate chapter. The third chapter is reserved to explain our approach for emotional body gesture recognition using the part-based model. we proposed a hybrid approach that incorporates two techniques which are: Motion estimation technique OF and temporal normalization method HMI to obtain a motion representation of hand that will be the input of deep stacked auto-encoder to classify hand gestures. Then, we took into consideration the relation between hand and face and we suggested a deep spatio-temporal approach based on the concatenation of temporal normalization method EBMI and deep learning method SSAE to classify emotions. In this part, we demonstrated that deep learning techniques outperform traditional machine learning techniques.

In the fourth chapter, we explained our proposed approach using the kinematic-based model. This model was used to precise exactly the position of body joints that leads to the good detection and tracking of human skeleton. We combined pose estimation and emotion classification to propose a new deep multi-stage architecture able to deal with both tasks by exploiting the strong points of CNN models pre-trained. The aim of this part is to improve the 2D human pose representation through using new structure of skeleton features based on transfer learning model DCN rather than using the skeleton information obtained from the Kinect tool. This 2D pose representation called HPI will be the input of the CNN model based on VGG architecture to classify emotions. In this part, we demonstrated that transfer learning techniques outperform traditional machine learning techniques.

The fifth chapter was devoted to present our proposed approach for speech emotion recognition. Speech is the fastest normal way to communicate amongst human. This reality motivated us to identify the emotional conditions of the person utterer by utilizing his/her voice automatically. We proposed a deep temporal-cepstrum representation based on the concatenation of spectral features, temporal derivatives features, and deep learning classifier for speech emotion recognition.

In general, the results obtained for both modalities, body gesture and speech, using our suggested approaches are very promising and competitive over existing methods in the state of the art. Our dissertation shows the feasibility of using automatic extracted cues to analyze the psychological states as an attractive alternative to manual annotations of behavioral cues.

The work performed has been described as fully and clearly as possible. As future work, we aim to test the robustness of our proposed approaches using large datasets. In fact, it is very difficult to distinguish emotions between only few individuals and the differences are not clear with a small set of data. Moreover, directed bodily action tasks differ in appearance and timing from affective and spontaneous body gesture characterized by its freedom and liberty. Therefore, data recording and processing for the last one is challenging, and many restrictions exist that make the identification and interpretation of spontaneous body gesture more difficult in comparison to other modalities. So, it is possible to merge the first sub-system which is body gesture with the second one which is speech by applying a technique of fusion for the good interpretation of the different psychological states. Our desired objective in the future is to design a non-verbal behavior analyzer tool that could be helpful in some domains like e-learning, e-therapy, enhanced websites customization, entertainment, job interview, etc.

Bibliography

[1] Bulat A. and Tzimiropoulos G. Human pose estimation via convolutional part heatmap regression. In *European Conference on Computer Vision (ECCV)*, pages 717–732, 2016.

[2] Choudhury A., Talukdar A. K., and Sarma K. K. A novel hand segmentation method for multiple-hand gesture recognition system under complex background. In *International Conference on Signal Processing and Integrated Networks (SPIN)*, pages 136–140, 2014.

[3] Hans A. and Hans E. Kinesics, haptics and proxemics: Aspects of non-verbal communication. *IOSR Journal of Humanities and Social Science (IOSR-JHSS)*, 20(2):47–52, 2015.

[4] Kapur A., Kapur A., Virji-Babul N., Tzanetakis G., and Driessen P. Gesture-based affective computing on motion capture data. In *International Conference on Affective Computing and Intelligent Interaction*, pages 1–7, 2005.

[5] Kolakowska A., Landowska A., Szwoch M., Szwoch W., and Wrobel M. R. Modeling emotions for affect-aware applications. *ESENCJA*, 55, 2015.

[6] Krizhevsky A., Sutskever I., and Hinton G. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[7] Marcos-Ramiro A., Pizarro-Perez D., Marron-Romera M., Nguyen L., and Gatica-Perez D. Body communicative cue extraction for conversational analysis. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 2013.

[8] Mehrabian A. Communication without words. *Psychology today*, 2(4), 1968.

[9] Toshev A. and Szegedy C. Deeppose: Human pose estimation via deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA*, pages 1653–1660, 2014.

[10] Toshev A. and Szegedy C. Coco 2018 keypoint detection task. In *https://cocodataset.org/#keypoints-2018*, 2018.

[11] Voulodimos A., Doulamis N., Doulamis A., and Protopapadakis E. Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci*, pages 1–13, 2018.

[12] Yilmaz A., Javed O., and Shah M. Object tracking: A survey. In *ACM Computing Surveys*, volume 38, 2006.

[13] Zamil A. A. A., Hasan S., Jannatul Baki, S. MD., Jawad A. MD., and Isra Z. Emotion detection from speech signals using voting mechanism on classified frames. In *2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST)*, pages 281–285. IEEE, 2019.

[14] Ingale A. B. and Chaudhari D. Speech emotion recognition using hidden markov model and support vector machine. *Journal of Advanced Engineering Research and Studies*, 1(3):316–318, 2012.

[15] Huffcut A. I., Conway J. M., Roth P. L., and Stone N. J. Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, 86(5):897–913, 2001.

[16] Ahad A. R., Tan Md., Kim J. K., and Ishikawa S. Motion history image: Its variants and applications. *Journal of Machine Vision and Applications*, 23:255–281, 2012.

[17] Imada A. S. and Hakel M. D. Influence of nonverbal communication and rater proximity on impressions and decisions in simulated employment interviews. *Journal of Applied Psychology*, 62(3):295–300, 1977.

[18] Kumthekar A. V. and Patil J. K. Key frame extraction using color histogram method. *International Journal of Scientific Research Engineering and Technology (IJSRET)*, 2:207–214, 2013.

[19] Abdul Malik B., Nasir R., Noor U., Jamil A., Khan M., Mi Young L., Soonil K., and Sung Wook B. Deep features-based speech emotion recognition for smart affective services. *Multimedia Tools and Applications*, 78(5):5571–5589, 2019.

[20] Basharirad B. and Moradhaseli M. Speech emotion recognition methods: A literature review. In *International Conference on Applied Science and Technology*, 2017.

[21] De Gelder B. Emotions and the body. In *Oxford University Press, New York*, 2016.

[22] Gelder B. Why bodies? twelve reasons for including bodily expressions in affective neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364:3475–3484, 2009.

[23] Wrede B. and Shriberg E. Spotting "hot spots" in meetings: Human judgments and prosodic cues. In *Proceedings of Euro speech*, 2003.

[24] Hanson B. A. and Applebaum T. H. Robust speaker-independent word recognition using static, dynamic and acceleration features: Experiments with lombard and noisy speech. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 857–860. IEEE, 1990.

[25] Bedregal B. R. C., Dimuro G. P., and Costa A. C. R. Hand gesture recognition in an interval fuzzy approach. *Tema- Trends in computational and applied mathematics*, 8(1):21–31, 2007.

[26] Buzzelli M. Mazzini D. Bianco, S. and R Schettini. A fully convolutional network for salient object detection. In *International Conference on Image Analysis and Processing*, page 82â“92, 2017.

[27] Wang C., Liu Z., and Chan S. Superpixel-based hand gesture recognition with kinect depth camera. *IEEE transactions on multimedia*, pages 29–39, 2015.

[28] Lee C. C., Mower E., Busso C., Lee S., and Narayanan S. Emotion recognition using a hierarchical binary decision tree approach. In *Interspeech*, pages 320–323, 2009.

[29] Izard C. E. Human emotions. In *New York: Plenum Press*, 1977.

[30] Parsons C. K. and Liden R. C. Interviewer perceptions of applicant qualifications: A multivariate field study of demographic characteristics and nonverbal cues. *Journal of Applied Psychology*, 69(4):557–568, 1984.

[31] Lisetti C. L. and Nasoz F. Using noninvasive wearable computers to recognize human emotions from physiological signals. *Journal on applied Signal Processing*, pages 1672–1687, 2004.

[32] McNeill D. Gesture and thought. In *the university of Chicago Press books*, 2005.

[33] Sanchez-Cortes D., Aran O., Schmid Mast M., and Gatica-Perez D. A nonverbal behavior approach to identify emergent leaders in small groups. *IEEE Transactions on Multimedia*, 4(3):816–832, 2012.

[34] Yu D. and Li D. Automatic speech recognition: A deep learning approach. *Signals and Communication Technology*, 2015.

[35] Jayagopi D. B., Hung H., Yeo C., and Gatica-Perez D. Modeling dominance in group conversations using nonverbal activity cues. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(3):501–513, 2009.

[36] Andrius Dzedzickis, Arturas Kaklauskas, and Vytautas Bucinskas. Human emotion recognition: Review of sensors and methods. *Sensors*, 20(3), 2020.

[37] Couzon E. and Dorn F. Les emotions: developper son intelligence emotionnelle. *Issyles-Moulineaux: ESF editeur*, 2009.

[38] Stergiopoulou E., Sgouropoulos K., Nikolaou N., Papamarkos N., and Mitianoudis N. Real time hand detection in a complex background. *Eng. Appl. Artif. Intell*, 35:54–70, 2014.

[39] Ekman and Friesen. The repertoire of nonverbal behavior: Categories, origins, and coding. *Journal of the International Association for Semiotic Studies*, 1:49–98, 2009.

[40] Baradel F., Wolf C., and Mille J. Pose-conditioned spatio-temporal attention for human action recognition. In *arXiv preprint arXiv:1703.10106*, 2017.

[41] Burkhardt F., Paeschke A., Rolfes M., Walter S., and Benjamin W. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*, 2005.

[42] Noroozi F., Corneanu C. A., Kaminska D., Sapinski T., Escalera S., and Anbarjafari G. Survey on emotional body gesture recognition. *IEEE Transactions on Affective Computing arXiv preprint arXiv:1801.07481*, 2018.

[43] Pianesi F., Mana N., Cappelletti A., Lepri B., and Zancanaro M. Multimodal recognition of personality traits in social interactions. In *International Conference on Multimodal Interfaces*, 2008.

[44] Castellano G., Kessous L., and Caridakis G. Emotion recognition through multiple modalities: face, body gesture, speech. *Affect and emotion in human-computer interaction*, pages 92–103, 2008.

[45] Castellano G., Villalba S., and Camurri A. Recognising human emotions from body movement and gesture dynamics. In *International Conference on Affective Computing and Intelligent Interaction*, pages 71–82, 2007.

[46] Deshmukh G., Gaonkar A., Golwalkar G., and et al. Speech based emotion recognition using machine learning. In *International Conference on Computing Methodologies and Communication (ICCMC)*, pages 812–817, 2019.

[47] Howard G., Zhu M., Chen B., Kalenichenko D., Wang W., Weyand T., Andreetto M., and Adam H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. In *arXiv preprint arXiv:1704.04861*, 2017.

[48] Yu G., Eric P., Hai-Xiang L., and Jaap van den H. Speech emotion recognition using voiced segment selection algorithm. In *Proceedings of the Twenty-second European Conference on Artificial Intelligence*, pages 1682–1683. IOS Press, 2016.

[49] Sree G. D., Chandrasekhar P., and Venkatesshulu B. Svm based speech emotion recognition compared with gmm-ubm and nn. *Journal of Engineering Science and Computing (IJESC)*, 6(11):3293–3298, 2016.

[50] Todor Ganchev, Nikos Fakotakis, and George Kokkinakis. Comparative evaluation of various mfcc implementations on the speaker verification task. In *Proceedings of the SPECOM*, volume 1, pages 191–194, 2005.

[51] Yuanbo Gao, Baobin Li, Ning Wang, and Tingshao Zhu. Speech emotion recognition using local and global features. In *International Conference on Brain Informatics*, pages 3–13. Springer, 2017.

[52] Murthy G.R.S. and Jadon R. S. A review of vision based hand gestures recognition. *International journal of Information Technology and Knowledge Management*, 2:405–410, 2009.

[53] Alshamsi H., Kepuska V., Alshamsi H., and Meng H. Automated speech emotion recognition on smart phones. In *Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, 2018.

[54] Coskun H. Human pose estimation with cnns and lstms. In *Master's Thesis, Technical University of Munich Germany*, 2016.

[55] Gunes H. and Piccardi M. The fabo database. In *http://mmv.eecs.qmul.ac.uk/fabo/*, 2005.

[56] Gunes H. and Piccardi M. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *The 18th International Conference on Pattern Recognition (ICPR), Hong Kong, China.* IEEE, 2006.

[57] Gunes H. and Piccardi M. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30:1334–345, 2007.

[58] Gunes H. and Piccardi M. Automatic temporal segment detection and affect recognition from face and body display. *IEEE Transaction on Systems, Man and Cybernetics*, 39:64–4, 2009.

[59] Kaur H. and Rani J. A review: Study of various techniques of hand gesture recognition. In *IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*, pages 1–5, 2016.

[60] Liang H., Zhao Y., Wei J., Quan D., Cheng R., and Wei Y. Robust hand detection and tracking based on monocular vision. In *IEEE International Conference on Intelligent Human-Machine Systems and Cybernetics*, 2014.

[61] Liang H., Zhao Y., Wei J., Quan D., Cheng R., and Wei Y. Robust hand detection and tracking based on monocular vision. In *IEEE International Conference on Intelligent Human-Machine Systems and Cybernetics*, 2014.

[62] Md Rashidul Hasan, Mustafa Jamil, MGRMS Rahman, et al. Speaker identification using mel frequency cepstral coefficients. *variations*, 1(4), 2004.

[63] Jemel I., Ejbali R., and Zaied M. Computer control system using a virtual keyboard. In *International Conference on Machine Vision*, 2015.

[64] khalifa I., Ejbali R., and Zaied M. Hand motion modeling for psychology analysis in job interview using optical flow-history motion image (of-hmi). In *The 10th International Conference on Machine Vision ICMV*, 2017.

[65] khalifa I., Ejbali R., and Zaied M. Body gesture modeling for psychology analysis in job interview based on deep spatio-temporal approach. In *Parallel and Distributed Computing, Applications and Technologies*, volume 931, pages 274–284. Springer, 2018.

[66] Prakash J. and Gautam U. K. Hand gesture recognition. *Int. J. Recent Technol. Eng*, 7:54–59, 2019.

[67] Rong J., Li G., and Chen Y.-P. P. Acoustic feature selection for automatic emotion recognition from speech. *Process Manag*, 45(3):315–328, 2009.

[68] Russell J. A circumplex model of affect. *Journal of personality and social Psychology*, 39, 1980.

[69] Russell J. and Mehrabian A. Evidence for a three-factor theory of emotions. *J. Research in Personality*, 11:273–294, 1977.

[70] Wang J., Ma Y., Zhang L., and Gao RX. Deep learning for smart manufacturing: Methods and applications. *J Manuf Syst*, 48:144–156, 2018.

[71] Yeh J.-H., Pao T.-L., Lin C.-Y., Tsai Y.-W., and Chen Y.-T. Segment-based emotion recognition from continuous mandarin chinese speech. *Comput. Human Behav*, 27(5):1545–1552, 2011.

[72] Wachs J. P., Kolsch M., Stern H., and Edan Y. Vision-based hand-gesture applications. In *Commun. ACM*, volume 54, pages 60–71, 2011.

[73] Sonkusare J. S., Chopade N. B., Sor R., and Tade S. L. A review on hand gesture recognition system. In *International Conference on Computing Communication Control and Automation*, page 790â“794, 2015.

[74] Burgoon JK., Jensen ML., Meservy TO., Kruse J., and Nunamaker JF. Augmenting human identification of emotional states in video. In *International conference on intelligent data analysis*, 2005.

[75] Pansare J.R., Gawande S.H., and Ingle M. Real-time static hand gesture recognition for american sign language (asl) in complex background. *Journal of Signal and Information Processing*, 3:364–367, 2012.

[76] He K., Zhang X., Ren S., and Sun J. Deep residual learning for image recognition. In *arXiv preprint arXiv:1512.03385*, 2015.

[77] Mantripragada K., Trigo F. C., Martins FP. R., and Fleury A. T. Vehicle tracking using feature matching and kalman filtering. In *ABCM Symposium Series in Mechatronic*, 2014.

[78] Simonyan K. and Zisserman A. Very deep convolutional networks for large-scale image recognition. In *arXiv preprint arXiv:1409.1556*, 2014.

[79] Scherer K. R. and Moors A. The emotion process: Event appraisal andcomponent differentiation. *Annual Review of Psychology*, 70:719–745, 2019.

[80] Asifullah Khan, Anabia Sohail, Umme Zahoora, and Aqsa Saeed Qureshi. A survey of the recent architectures of deep convolutional neural networks. *arXiv preprint arXiv:1901.06032*, 2019.

[81] Bottou L. Large-scale machine learning with stochastic gradient descent. In *International Conference on Computational Statistics, Paris, France*, 2010.

[82] Chen L., Mao X., Xue Y., and et al. Speech emotion recognition: Features and classification models. *Digital Signal Processing*, 22(6):1154–1160, 2012.

[83] Dipietro L., Sabatini A. M., Member S., and Dario P. A survey of glove-based systems and their applications. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38:461–482, 2008.

[84] Kaiser L., Gomez A.N., and Chollet F. Depthwise separable convolutions for neural machine translation. In *arXiv preprint arXiv:1706.03059*, 2017.

[85] Perez L. and Wang J. The effectiveness of data augmentation in image classification using deep learning. In *arXiv preprint arXiv:1712.04621*, 2017.

[86] Aristomenis L. S. and Tsihrintzis G. A. Evaluation of mpeg-7 descriptors for speech emotional recognition. In *2012 Eighth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 98–101. IEEE, 2012.

[87] Li and Yi. Hand gesture recognition using kinect. In *International Conference on computer Science and Automation Engineering*, 2012.

[88] Steven R Livingstone and Frank A Russo. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5), 2018.

[89] Chaudhari M., Sondur S., and Vanjare G. A review on face detection and study of viola jones method. *International Journal of Computer Trends and Technology (IJCTT)*, 25(1):54–61, 2015.

[90] Coulson M. Attributing emotion to static body postures: recognition accuracy, confusions, and viewpoint dependence. *Journal of Nonverbal Behavior*, pages 39–117, 1992.

[91] ElAyadi M., Kamel M. S., and Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572â“587, 2011.

[92] Greenwald M., Cook E., and Lang P. Affective judgment and psycho physiological response: Dimensional covariation in the evaluation of pictorial stimuli. *J. Psychophysiology*, 3:51–64, 1989.

[93] Grimm M., Kroschel K., Mower E., and Narayanan S. Primitives-based evaluation and estimation of emotions in speech. *Speech Commun*, 49(10):787â“800, 2007.

[94] Hassan M., Ahmad T., and Javaid M.A. Human activity recognition using motion history algorithm. *International Journal of Scientific and Engineering Research*, 5:1037–1044, 2014.

[95] Kipp M. Gesture generation by imitation from human behavior to computer character animation. In *Boca Raton, Florida USA, Dissertation.com*, 2004.

[96] Oudah M., Abdulelah Al-Naji A., and Chahl J. Hand gesture recognition based on computer vision: A review of techniques. *MDPI: Journal of Imaging*, 2020.

[97] Paleari M., Benmokhtar R., and Huet B. Evidence theory-based multimodal emotion recognition. In *Springer-Verlag Berlin Heidelberg,(Eds.):MMM, LNCS 5371*, 2009.

[98] Pantic M., Sebe N., Cohn J. F., and Huang T. Affective multimodal human-computer interaction. In *International Conference on Multimedia*, pages 669–676. ACM, 2005.

[99] Patel M. and Bhatt B. A comparative study of object tracking techniques. *International Journal of Innovative Research in Science, Engineering and Technology*, 4:1361–1364, 2015.

[100] Sakkari M., Zaied M., and Amar C. B. Using hand as support to insert virtual object in augmented reality applications. *Journal of Data Processing*, 2:10–24, 2012.

[101] Sara M., Saeed S., and Azam R. Speech emotion recognition based on a modified brain emotional learning model. *Biologically inspired cognitive architectures*, 19:32–38, 2017.

[102] Van den Bergh M., Carton D., De Nijs R., Mitsou N., Landsiedel C., Kuehnlenz K., Wollherr D., Van Gool L., and Buss M. Real-time 3d hand gesture interaction with a robot for understanding directions from humans. In *Proceedings of the 2011 Ro-Man, Atlanta, GA, USA*, pages 357–362, 2011.

[103] P. Gehler M. Andriluka, L. Pishchulin and S. Bernt. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[104] Sayf A Majeed, Hafizah Husain, Salina Abdul Samad, and Tariq F Idbeaa. Mel frequency cepstral coefficients (mfcc) feature extraction enhancement in the application of speech recognition: A comparison study. *Journal of Theoretical & Applied Information Technology*, 79(1), 2015.

[105] Brian Mc., Raffel C., Liang D., PW Ellis D., McVicar M., Battenberg E., and Nieto O. librosa: Audio and music signal analysis in python. In *14th python in science conference*, pages 18–25, 2015.

[106] Anderson N. and Shackleton V. Decision making in the graduate selection interview: Afield study. *Journal of Occupational Psychology*, 63(1):63–76, 1990.

[107] Jammalamadaka N., Zisserman A., Eichner M., Ferrari V., and Jawahar C. V. Human pose evaluator dataset. In *http://www.robots.ox.ac.uk/ vgg/data/poseevaluation/*.

[108] Kolotouros N., Pavlakos G., Black M. J., and Danilidis K. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *International Conference on Computer Vision*, pages 2252–2261, 2019.

[109] Fatemeh Noroozi, Tomasz Sapiński, Dorota Kamińska, and Gholamreza Anbarjafari. Vocal-based emotion recognition using random forests and decision tree. *International Journal of Speech Technology*, 20(2):239–246, 2017.

[110] Barros P., Jirak D., Weber C., and Wermter S. Emotion-modulated attention improves expression recognition. *Journal of Neural Networks*, 72:140–151, 2015.

[111] Barros P., Parisi G., Weber C., and Wermter S. Emotion recognition via body gesture : A deep learning model. *Journal of NeuroComputing*, pages 104–114, 2017.

[112] Ekman P. Universal and cultural differences in facial expression of emotion. *Sym. Motiv*, 19:207–283, 1971.

[113] Ekman P. Strong evidence for universals in facial expressions: A reply to russell's mistaken critique. *Psychol. Bull*, 115(2):268–287, 1994.

[114] Ekman P., Friesenet W.V., and Ellsworth P. What emotion categories or dimensions can observers judge from facial behavior. In *Emotion in the human face. New York : Cambridge University Press*, 1982.

[115] Garg P., Aggarwal N., and Sofat S. Vision based hand gesture recognition. *Journal of Computer, Electrical, Automation, Control and Information Engineering*, 3:186–191, 2009.

[116] Sathit P. Improvement of speech emotion recognition with neural network classifier by using speech spectrogram. In *2015 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 73–76. IEEE, 2015.

[117] Shegokar P. and Sircar P. Continuous wavelet transform based speech emotion recognition. In *2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS)*, pages 1–8. IEEE, 2016.

[118] Vincent P., Larochelle H., Bengio Y., and Manzagol P. A. Extracting and composing robust features with denoising autoencoders. In *International Conference on Machine Learning, Helsinki, Finland*, 2008.

[119] Jack Parry, Dimitri P., Georgia C., Pauline L., Rebecca M., Michael B., and Gregor H. Analysis of deep learning architectures for cross-corpus speech emotion recognition. *Proc. Interspeech 2019*, pages 1656–1660, 2019.

[120] Anastasiya S Popova, Alexandr G R., and Alexander A P. Emotion recognition in sound. Springer.

[121] Kishore P.V.V. and Prasad M.V.D. Optical flow hand tracking and active contour hand shape features for continuous sign language recognition with artificial neural networks. *International Journal of Software Engineering and Its Applications*, 9:231–250, 2015.

[122] Chen Q., Georganas N.D., and Petriu E. M. Real-time vision-based hand gesture recognition using haar-like features. In *Proceedings of the 2007 IEEE instrumentation measurement technology conference IMTC*, pages 1–6, 2007.

[123] Ejbali R., Zaied M., and Amar C. B. Computer control system using a virtual keyboard. In *International Conference on Machine Vision*, 2014.

[124] Plutchik R. A general psycho-evolutionary theory of emotion. 1980.

[125] Plutchik R. 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*, 89(4):344–350, 2001.

[126] Suriya R. and Vijayachamundeeswari V. A survey on hand gesture recognition for simple mouse control. In *International Conference on Information Communication and Embedded Systems(ICICES)*, pages 1–5, 2014.

[127] Posthuma R. A., Morgeson F. P., and Campion M. A. Beyond employment interview validity: A comprehensive narrative review of recent research and trends over time. *Journal of Personnel Psychology*, 55(1), 2002.

[128] Forbes R. J. and Jackson P. R. Non-verbal behaviour and the outcome of selection interviews. *Journal of Occupational Psychology*, 53(1):65–72, 1980.

[129] Rautaray, Siddharth S., and Agrawal A. Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, pages 1–54, 2012.

[130] Wang R.Y. and Popovic J. Real-time hand-tracking with a color glove. In *ACM SIGGRAPH*, volume 28, pages 1–8, 2009.

[131] Abrilian S. Representation de comportements emotionnels multimodaux spontanes : Perception, annotation et synthese. In *These en informatique de l'Universite Paris*, 2007.

[132] Athavale S. and Deshmukh M. Dynamic hand gesture recognition for human computer interaction; a comparative study. *International Journal of Engineering Research and General Science*, 2:38–55, 2016.

[133] Bianco S., Buzzelli M., Mazzini D., and Schettini R. Logo recognition using cnn features. In *Image Analysis and Processing ICIAP*, pages 438–448, 2015.

[134] Bianco S., Buzzelli M., Mazzini D., and Schettini R. Deep learning for smart manufacturing: Methods and applications. *Neurocomputing*, 245:23–30, 2017.

[135] Chen S., Tian Y., Liu Q., and Metaxas D. N. Recognizing expressions from face and body gesture by temporal normalized motion and appearance features. *Journal of Image and Vision Computing*, 3:175–185, 2013.

[136] Crampton S. and Betke M. Counting fingers in real time: A webcam-based human-computer interface with game applications. In *Universal access in human-computer interaction*, 2003.

[137] Desai S. and Desai A. A. human computer interaction through hand gestures for home automation using microsoft kinect. In *International Conference on Communication and Networks*, pages 19–29, 2017.

[138] Gupta S., Bagga S., and Sharma D. K. Hand gesture recognition for human computer interaction and its applications in virtual reality. *In: Gupta D., Hassanien A., Khanna A. (eds) Advanced Computational Intelligence Techniques for Virtual Reality in Healthcare. Studies in Computational Intelligence*, 875, 2020.

[139] Hassairi S., Ejbali R., and Zaied M. Sparse wavelet auto-encoders for image classification. In *International Conference on Digital Image Computing: Techniques and Applications*, pages 1–6. IEEE, 2016.

[140] Hassairi S., Ejbali R., and Zaied M. A deep stacked wavelet auto-encoders to supervised-feature extraction to pattern classification. *Multimedia Tools and Applications*, 2017.

[141] Mehdi S. and Khan Y. Sign language recognition using sensor gloves. In *9th international conference on neural information processing*, volume 5, pages 2204–2206, 2002.

[142] Mitra S. and Acharya T. Gesture recognition: A survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 37:311–324, 2007.

[143] Nilu S., RA K., and Raj S. Mfcc and prosodic feature extraction techniques: A comparative study. *International Journal of Computer Applications*, 54(1), 2012.

[144] Piana S., Stagliano A., Odone F., verri A., and Camurri A. Real-time automatic emotion recognition from body gestures. *arXiv preprint arXiv:1402.5047*, 2014.

[145] Saha S., Datta S., Konar A., and R. Janarthanan. A study on emotion recognition from body gestures using kinect sensor. In *Communications and Signal Processing (ICCSP)*, pages 56–60, 2014.

[146] Shantaiya S., Verma K., and Mehta K. Multiple object tracking using kalman filter and optical flow. *European Journal of Advances in Engineering and Technology*, 2:34–39, 2015.

[147] Thai Ly S., Lee G. S., Kim S. H., and Yang H. J. Emotion recognition via body gesture: Deep learning model coupled with keyframe selection. In *International Conference on Machine Learning and Machine Intelligence*, pages 27–31, 2018.

[148] Tomkins S. Affect theory. In *In K. R. Scherer, P. Ekman (Eds.) Approaches to emotion, Hillsdale, NJ : Erlbaum*, 1984.

[149] Wu S., Falk T. H., and Chan W.-Y. Automatic speech emotion recognition using modulation spectral features. *Speech Commun*, 53(5):768â“785, 2011.

[150] Zagbani S., Jaouedi N., Boujnah N., and Bouhlel M. S. Real-time hand gesture recognition based on feature points extraction. In *International Conference on Machine Vision*, 2016.

[151] Baltrusaitis T., McDuff D., Banda N., Mahmoud M., El Kaliouby R., Robinson P., and Picard R. Real-time inference of mental states from facial expressions and upper body gestures. In *Automatic Face and Gesture Recognition and Workshops (FG2011), USA*, pages 909–914, 2011.

[152] Bouchrika T., Zaied M., Jemai O., and Amar CB. Neural solutions to interact with computers by hand gesture recognition. *Multimedia tools and applications*, 72:2949–2975, 2014.

[153] Simon T., Joo H., Matthews I., and Sheikh Y. Hand keypoint detection in single images using multiview bootstrapping. In *IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017.

[154] Vogt T., Andre E., and Wagner J. Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realization. In *Affect and emotion in human-computer interaction*, pages 75 –91, 2008.

[155] Roy Tanmoy, Chakraverty S., Marwala T., and Satyakama P. Introducing new feature set based on wavelets for speech emotion classification. In *2018 IEEE Applied Signal Processing Conference (ASPCON)*, pages 124–128. IEEE, 2018.

[156] Martin V. and Robert V. Recognition of emotions in german speech using gaussian mixture models. In *Multimodal Signals: Cognitive and Algorithmic Issues*, pages 256–263. Springer, 2009.

[157] Kulkarni V. S. and Lokhande S. D. Appearance based recognition of american sign language using gesture segmentation. *Int. J. Comput. Sci. Eng*, 2:560â“565, 2010.

[158] Dai W., Han D., Dai Y., and Xu D. Emotion recognition and affective computing on vocal social media. In *Information management*, 2015.

[159] Wang W., Enescu V., and Sahli H. Adaptive real-time emotion recognition from body movements. *Transactions on Interactive Intelligent Systems*, 5(4), 2015.

[160] Baveye Y., Dellandrea E., Chamaret C., and Chen L. Liris-accede: Avideo database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55, 2015.

[161] Chen Y., Luo B., Chen Y.-L., Liang G., and Wu X. A real-time dynamic hand gesture recognition system using kinect sensor. In *International Conference on Robotics and Biomimetics (ROBIO)*, pages 2026–2030, 2015.

[162] Fang Y., Cheng J., Wang K., and Lu H. Hand gesture recognition using fast multi-scale analysis. In *International conference on Image and graphics*, volume 5, pages 694–698, 2007.

[163] Fang Y., Wang K., Cheng J., and Lu H. A real-time hand gesture recognition method. In *IEEE International Conference on Multimedia and Expo*, page 995â“998, 2007.

[164] Guo Y., Zhao Z., Ma Y., and et al. Speech augmentation via speaker-specific noise in unseen environment. In *Interspeech*, pages 1781–1785, 2019.

[165] Shi Y. Tensorflow 1 detection model zoo. In *https://github.com/tensorflow/models/blob/master/research/object_detection/g3doc/tf1_detection_zoo.md*, Accessed: 05/06/2019.

[166] Shi Y., Yang H., Gong M., Liu X., and Xia Y. Fast and robust key frame extraction method for video copyright protection. *Journal of Electrical and Computer Engineering*, 2017.

[167] Cao Z., Hidalgo G., Simon T., Wei S. E., and Sheikh Y. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):172–186, 2019.

[168] Chen Z., Kim J., Liang J., Zhang J., and Yuan Y. B. Real-time hand gesture recognition using finger segmentation. *Hindawi Publishing Corporation e Scientific World Journal*, 2014.

[169] Witkower Z. and Tracy J. Bodily communication of emotion: Evidence for extra facial behavioral expressions and available coding systems. *Journal of Emotion Review*, 11:184–193, 2018.

[170] Liu Z-T., Wu M., and Cao W-H. Speech emotion recognition based on feature selection and extreme learning machine decision tree. *Neurocomputing*, 273:271–280, 2018.