OULU 2022
C 837

ACTA UNIVERSITATIS OULUENSIS

Wheidima Carneiro de Melo

DEEP REPRESENTATION LEARNING FOR AUTOMATIC DEPRESSION DETECTION FROM FACIAL EXPRESSIONS

UNIVERSITY OF OULU GRADUATE SCHOOL;
UNIVERSITY OF OULU,
FACULTY OF INFORMATION TECHNOLOGY AND ELECTRICAL ENGINEERING

ACTA UNIVERSITATIS OULUENSIS
C Technica 837

WHEIDIMA CARNEIRO DE MELO

DEEP REPRESENTATION LEARNING FOR AUTOMATIC DEPRESSION DETECTION FROM FACIAL EXPRESSIONS

Academic dissertation to be presented with the assent of the Doctoral Programme Committee of Information Technology and Electrical Engineering of the University of Oulu for public defence in the OP auditorium (L10), Linnanmaa, on 12 August 2022, at 12 noon

UNIVERSITY OF OULU, OULU 2022

Acta Univ. Oul. C 837, 2022

Supervised by
Associate Professor Miguel Bordallo López

Reviewed by
Associate Professor Carlos Roberto del Blanco
Professor Alexandros Iosifidis

Opponent
Docent Heikki Huttunen

ISBN 978-952-62-3366-6 (Paperback)
ISBN 978-952-62-3367-3 (PDF)

ISSN 0355-3213 (Printed)
ISSN 1796-2226 (Online)

Cover Design
Raimo Ahonen

PUNAMUSTA
TAMPERE 2022

Carneiro de Melo, Wheidima, Deep representation learning for automatic depression detection from facial expressions.

University of Oulu Graduate School; University of Oulu, Faculty of Information Technology and Electrical Engineering

Acta Univ. Oul. C 837, 2022

University of Oulu, P.O. Box 8000, FI-90014 University of Oulu, Finland

Abstract

Depression is a prevalent mental disorder that severely affects an individual's quality of life. Traditional diagnostic methods rely on either clinician's evaluation of symptoms reported by an individual or self-report instruments. These subjective assessments have resulted in difficulties to recognize depression. This scenario has motivated the development of automatic diagnostic systems to provide objective and reliable information about depressive states. Recently, a growing interest has been generated in developing such systems based on facial information since there exists evidence that facial expressions convey valuable information about depression.

This thesis proposes computational models to explore the correlations between facial expressions and depressive states. Such exploration is a challenging task because 1) the difference in facial expressions along different depression levels may be small and 2) the complexities involved in facial analysis. From this perspective, we investigate different deep learning techniques to effectively model facial expressions for automatic depression detection. Specifically, we design architectures that model the appearance and dynamics of facial videos. For that, we analyze structures that explore either a fixed or multiple spatiotemporal ranges. Our findings suggest that the use of a structure with multiscale feature extraction ability contributes to learning depression representation. We also demonstrate that depression distributions increase the robustness of depression estimations.

Another key challenge in this application is the scarcity of labelled data. This limitation leads to the need of efficient representation learning methods. To this end, we first develop a pooling method to encode facial dynamics into an image map, which may be explored by less complex deep models. In addition, we design an architecture to capture different facial expression variations by using a basic structure based on functions that explore features at multiple ranges without using trainable parameters. Finally, we develop an architecture to explore facial expressions related to depression and pain since depressed individuals may experience pain. To build this architecture, we use different strategies to efficiently extract multiscale features. Our experiments indicate that the proposed methods have the potential to generate discriminative representations.

Keywords: automatic depression detection, convolutional neural network, deep learning, facial expression analysis, representation learning

Carneiro de Melo, Wheidima, Piirteiden syväoppiminen masennuksen automaattiseen tunnistukseen kasvonilmeistä.
Oulun yliopiston tutkijakoulu; Oulun yliopisto, Tieto- ja sähkötekniikan tiedekunta
Acta Univ. Oul. C 837, 2022
Oulun yliopisto, PL 8000, 90014 Oulun yliopisto

Tiivistelmä

Masennus on yleinen mielenterveyden häiriö, joka heikentää merkittävästi yksilön elämänlaatua. Perinteiset diagnostiset menetelmät nojaavat joko kliinikon arvioon oireista potilaan kertomuksen perusteella tai itsearvioihin. Subjektiivinen arviointi on johtanut vaikeuksiin tunnistaa masennusta. Tämä motivoi kehittämään automaattisia diagnostiikkajärjestelmiä tarjoamaan objektiivista ja luotettavaa tietoa masennustiloista. Viime aikoina kiinnostus hyödyntää kasvoista saatavaa informaatiota kyseisissä järjestelmissä on noussut, sillä on pystytty osoittamaan ilmeiden välittävän arvokasta tietoa masennuksesta.

Tässä väitöskirjassa esitetään laskennallisia malleja tutkimaan korrelaatiota ilmeiden ja masennustilojen välillä. Tehtävä on haastava, sillä: 1) ilmeiden ja masennuksen eri tasojen väliset erot saattavat olla pieniä ja 2) kasvoanalyysiin liittyy monimutkaisuuksia. Tästä näkökulmasta tutkitaan erilaisia syväoppimistekniikoita mallintamaan tehokkaasti ilmeitä masennuksen automaattisessa tunnistuksessa. Erityisesti suunnitellaan arkkitehtuureja, jotka mallintavat kasvoja ja niiden dynamiikkaa videoista. Tätä varten analysoidaan rakenteita, jotka tutkivat kiinteää spatiotemporaalista aluetta sekä spatiotemporaalista moniskaalainformaatiota. Havaintojen pohjalta spatiotemporaalisten moniskaalarakenteiden käyttö parantaa piirteiden irrotuskykyä masennuksen esitystavan oppimisessa. Masennusjakaumien antama lisä masennusestimaattien luotettavuudessa osoitetaan.

Toinen tärkeä sovellushaaste on luokitellun datan niukkuus, mistä seuraa tarve oppia tehokkaita esitystapoja. Tätä varten aluksi kehitetään yhdistämismenetelmä kasvojen dynamiikan koodaamiseksi kuvakartalle, jota voidaan tutkia laskennallisesti kevyillä syväoppimismenetelmillä. Lisäksi suunnitellaan arkkitehtuuri, joka rekisteröi eri ilmeiden vaihtelua. Perusteena ovat funktiot, jotka tutkivat piirteitä useilla arvoalueilla ilman opittavia parametreja. Lopuksi kehitetään arkkitehtuuri tutkimaan masennukseen ja kivun tunteeseen liittyviä ilmeitä, sillä masentuneet ihmiset saattavat kokea kipua. Arkkitehtuurin rakentamisessa käytetään erilaisia strategioita spatiotemporaalisten moniskaalapiirteiden irrottamiseen. Laajat kokeet osoittavat, että esitetyillä menetelmillä on potentiaalia luoda erottelukykyisiä esitystapoja.

Asiasanat: automaattinen masennuksen tunnistaminen, esitystavan oppiminen, ilmeanalyysi, konvoluutioneuroverkko, syväoppiminen

To my beloved wife and daughter.

Acknowledgements

I had the privilege to carry out my research work in the Center for Machine Vision and Signal Analysis (CMVS), University of Oulu. During my journey, I was encouraged by many people, who helped me reach this milestone in my career. Firstly, I would like to thank my supervisor, Dr. Miguel Bordallo López, for his patience, valuable advice, and his help in different stages of my PhD. His guidance was essential to the development of the whole thesis. I will always be thankful for that. I would also like to thank Professor Olli Silvén for providing an excellent research environment and giving me the opportunity to conduct my studies at CMVS.

I would like to thank Professor Éric Granger (École de Technologie Supérieure, Canada) for his assistance and technical suggestions to improve my research work. I am also grateful to the members of my follow-up group, Dr. Li Liu and Dr. Xiaobai Li, for their insightful comments on my doctoral training plan. I shall also thank Professor Alexandros Iosifidis (Aarhus University, Denmark) and Professor Carlos Roberto del Blanco (Universidad Politécnica de Madrid, Spain) for their dedication in reviewing my thesis. I would also like to express gratitude to Professor Mourad Oussalah and Professor Guoying Zhao for their support. I am also grateful to my previous supervisors from Brazil, Professor Waldir Sabino da Silva Júnior and Professor Eddie Batista de Lima Filho, for their encouragement and help.

Additionally, I have to thank my colleagues at CMVS for sharing knowledge and having happy moments with me. It made my journey amazing and unforgettable. I would specially like to thank Usman Muhammad, Mohammad Tavakolian, and Tuomas Holmberg for treating me like a family member.

I want to express my special thanks to my family. I would not be able to finish this work without the unconditional support of my wife and daughter, and I can not express how grateful I am to my mother for all the sacrifices she made to support my studies during my entire life. Above all, I would like to thank God for this opportunity and all blessings.

Wheidima Carneiro de Melo
Montreal, Canada, May 2022

List of abbreviations

AAM	*Active Appearance Model*
AU	*Action Unit*
AVEC	*Audio-Visual Emotion Challenge*
BDI	*Beck Depression Inventory*
C3D	*Convolutional 3D*
CAM	*Class Activations Map*
CCA	*Canonical Correlation Analysis*
CNN	*Convolutional Neural Network*
CERT	*Computer Expression Recognition Toolbox*
DCT	*Discrete Cosine Transform*
DDDAMC	*Depressive Disorder Due to Another Medical Condition*
DMDD	*Disruptive Mood Dysregulation Disorder*
DMSN	*Decomposed Multiscale Spatiotemporal Network*
DSM-5	*Diagnostic and Statistical Manual of Mental Disorders*
DTL	*Deep Transformation Learning*
EEA	*European Economic Area*
EEG	*Electroencephalogram*
EOH	*Edge Orientation Histogram*
FACS	*Facial Action Coding System*
FDHH	*Feature Dynamic History Histogram*
FV	*Fisher Vector*
GAP	*Global Average Pooling*
GMHI	*Gabor Motion History Image*
HAMD	*Hamilton Rating Scale for Depression*
HCI	*Human-Computer Interaction*
HoT	*Histograms of Topographical*
I3D	*Inflated 3D*
LBP	*Local Binary Pattern*
LBP-TOP	*Local Binary Patterns on Three Orthogonal Planes*
LPQ	*Local Phase Quantization*
LSTM	*Long-Short Term Memory*
MAE	*Mean Absolute Error*
MDD	*Major Depressive Disorder*
MDN	*Maximization and Differentiation Network*

MHH	*Motion History Histogram*
MHI	*Motion History Image*
MSE	*Mean Squared Error*
MSN	*Multiscale Spatiotemporal Network*
OSDD	*Other Specified Depressive Disorder*
OSVR	*Ordinal Support Vector Regression*
PDD	*Premenstrual Dysphoric Disorder*
PHQ	*Patient Health Questionnaire*
RCNN	*Recurrent Convolutional Neural Network*
ReLU	*Rectified Linear Unit*
RMSE	*Root Mean Square Error*
RNN	*Recurrent Neural Network*
SCN	*Spatiotemporal Convolutional Network*
SVM	*Support Vector Machine*
S/M-IDD	*Substance/Medication-Induced Depressive Disorder*
UDD	*Unspecified Depressive Disorder*
WHO	*World Health Organization*

List of original publications

This thesis is based on the following articles, which are referred in the text by their Roman numerals (I–VI):

I Carneiro de Melo W, Granger E & Hadid A (2019) Combining global and local convolutional 3D networks for detecting depression from facial expressions. Proc. IEEE International Conference on Automatic Face & Gesture Recognition (FG), Lille, France, pp. 1–8.

II Carneiro de Melo W, Granger E & Hadid A (2019) Depression detection based on deep distribution learning. Proc. IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, pp. 4544–4548.

III Carneiro de Melo W, Granger E & Hadid A (2020) A deep multiscale spatiotemporal network for assessing depression from facial dynamics. IEEE Transactions on Affective Computing, doi:10.1109/TAFFC.2020.3021755.

IV Carneiro de Melo W, Granger E & Bordallo Lopez M (2020) Encoding temporal information for automatic depression recognition from facial analysis. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Barcelona, Spain, pp. 1080–1084.

V Carneiro de Melo W, Granger E & Bordallo Lopez M (2021) MDN: A deep maximization-differentiation network for spatio-temporal depression detection. IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2021.3072579.

VI Carneiro de Melo W, Granger E & Bordallo Lopez M (2022) Automatic facial expression analysis using decomposed multiscale spatiotemporal networks. Submitted for evaluation. ArXiv preprint arXiv:2203.11111, https://doi.org/10.48550/arXiv.2203.11111.

The author of the thesis is the first author in the aforementioned articles. The responsibilities as the first author include developing and implementing methods, conducting experiments, and writing manuscripts. The co-authors provided valuable suggestions about the experiments and assisted to improve the quality of the manuscripts.

Contents

1 Introduction

1.1 Background and motivation

Doctor: “How are you feeling?”
Patient: “I have been extremely sad for several days, feeling like crying all the time even when I have no reason to do that.”

Unfortunately, this answer reflects a common sign experienced by millions of people suffering from depression [1]. This disorder is associated with a negative condition of mind which remains for a long time. It is also characterized by a high tendency for recurrence, with around 60% of individuals who recover from a first episode experiencing a second episode [2]. This serious condition may interfere in different aspects of an individual’s life, such as productivity [3], and relationships [4, 5]. The increasing number of individuals with depression and immense economic burden has attracted the attention of authorities. It is estimated that the prevalence of such a mental health disorder across Europe is 6.38% [6], with higher incidence in women (7.7%) than in men (4.9%). The total economic costs of depression in the European Economic Area (EEA) in 2007 were €136.3 billion, where €99.3 billion were related to losses in work productivity and €37.0 billion were costs in health care [7].

A depressive state may have diverse manifestations and the categorization of these symptoms allows the definition of different depressive disorders. The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [8] of the American Psychiatric Association (APA) classifies depressive disorders into eight variants: Major Depressive Disorder (MDD), Persistent Depressive Disorder (Dysthymia), Disruptive Mood Dysregulation Disorder (DMDD), Premenstrual Dysphoric Disorder (PDD), Substance/Medication-Induced Depressive Disorder (S/M-IDD), Depressive Disorder Due to Another Medical Condition (DDDAMC), Other Specified Depressive Disorder (OSDD) or Unspecified Depressive Disorder (UDD). The most common and severe variant is MDD [9], also referred to as depression. The symptoms associated with depression may encompass alterations in appetite [10], fatigue [11], sleep disturbances [12], psychomotor retardation [13], reduced ability to concentrate [14], headache [15], backache [15], stomach ache [16], feelings of worthlessness or disproportionate guilt [8], anxiety [17], loss of pleasure and/or interest in persons or things [8]. In severe cases, depression leads to substance abuse [18] and suicidal behavior [19]. Studies [19, 20] show high prevalence of psychiatric disorders in suicidality, where depression is the most frequently reported. Furthermore, depression may elevate the chances of developing and

sometimes contribute to the progress of serious medical conditions, such as diabetes, cardiovascular disease, and cancer [21].

A correct clinical diagnosis of depression performed in a timely manner is crucial to recognize depression and start proper treatment so that patients achieve positive health outcomes. However, studies have shown that clinicians have difficulties to diagnose depression [22, 23]. Indeed, the evaluation of depression has a subjective nature since it depends on clinical interviews using self-report questionnaires (e.g., Patient Health Questionnaire (PHQ) [24]) or an assessment instrument like Hamilton Rating Scale for Depression (HAMD) [25] which is administrated by a physician. Inaccurate assessment of depression has resulted in a high number of false-positives that present serious consequences for the patients [23]. Moreover, although there exist effective treatments for depression, a study funded by the European Commission estimates that, in Europe, about 56% of patients suffering from depression receive no treatment [26]. This number is alarming because early treatment gives the patient the best chance of recovery. Among the reasons for this number are restricted or lack of accessibility to mental healthcare, payments for mental health services, and stigma against mental disorders [26].

The challenges in the evaluation of depression motivate the development of decision support diagnosis systems based on non-verbal behaviors. For instance, consider the case where the patient is suffering from depression, but he only reports physical symptoms. This would be a challenging scenario for a clinician, but the automatic system could analyze the non-verbal cues and alert the doctor that the patient may have depression. Another interesting case is when the patient is diagnosed with depression and starts his/her treatment. In this scenario, it would be very important to measure in short periods the depression level of the patient in order to verify the success of the treatment. Given the shortage of doctors, this seems impractical, but automatic systems could perform this verification favoring the whole clinical process. Consequently, such systems may contribute to the reliability and improvement of clinical assessment and monitoring by providing an accurate and objective estimation of depression levels. These solutions may also help to alleviate problems related to access and costs. For example, a system could be used in remote assessments, enabling an expansion of evaluations and assisting in the early diagnosis of depression, which may reduce the cost of treatment.

Automatic depression detection is an emerging field for automatic medical diagnosis that encompasses studies from psychology, medicine, affective computing, signal processing, computer vision, and computer science fields. It consists of different methods to automatically assess depression such as depression detection, depression severity recognition, and depression estimation. Depression detection refers to techniques that are used to identify depressed and non-depressed individuals. Depression severity

recognition denotes methods that output a discrete label related to depression severity. Depression estimation concerns techniques that generate a continuous depression score. The methods may analyze dependencies of depressive behaviors acquired by non-invasive sensors. Examples of these sensors are cameras, which capture facial expressions, body movements, gaze, and microphones, which record speech activities.

This thesis focuses on automatic depression detection from facial expressions. Diverse studies have found correlations between facial behavior and depressive states. Some of these studies show results that evidence diminished facial expressiveness in depressive individuals [27, 28, 29]. Findings have also associated depression with decreased positive emotional facial expressions [27, 30]. Other studies show that depressive patients present reduced smile intensity and shorter smile duration [31, 32] as well as less mouth movements [33, 9]. An increase in frowning has also been related to depression [34, 35]. Another common report during clinical interviews is the avoidance of eye contact with therapists by depressed patients [36, 37]. This set of facial cues has the potential to be automatically analyzed to support the assessment of depression by using representation methods which explore facial dependencies to learn depression patterns from videos, allowing the recognition of depression levels.

The work-flow of a typical automatic depression detection system based on facial expressions is exhibited in Fig. 1. Such systems are composed of three steps: preprocessing, feature extraction, and regression (or classification). The preprocessing step is responsible for detecting and cropping faces in the videos from depression datasets, allowing the subsequent steps to explore facial information. The feature extraction step is the essential component of these systems. It explores the correlations of facial expressions along different depression levels to produce descriptive and discriminative features. In other application domains, such as facial expression recognition [38], pain estimation [39], fatigue recognition [40], and autism detection [41], deep learning techniques have contributed to the progress in automatic facial analysis. These methods have the potential to generate effective representations by modeling multiple levels of abstraction in data [42]. Particularly, Convolutional Neural Networks (CNNs) have been considered a powerful class of deep models for learning visual representations. Finally, the regression step analyzes the representations and produces a depression score, or a classification stage estimates either the presence/absence of depression or its severity.

Although facial expressions convey rich information about depressive states and deep learning methods have the ability to obtain good representations for facial analysis, there are key challenges in developing automatic depression detection systems that explore facial expressions. In what follows, the main challenges to automatically analyze facial expressions to infer depression levels are described.

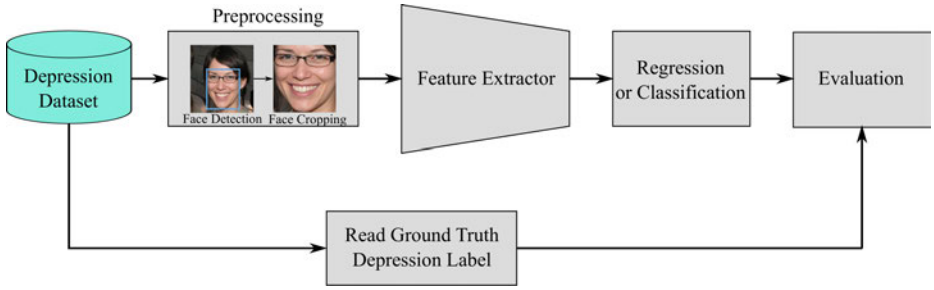


Fig. 1. The general work-flow of an automatic depression detection system using facial expressions.

Accessibility to depression data. The mental health data of a patient is protected by law. Collecting such sensitive data along with raw facial images is a difficult process due to privacy and data protection issues. Perhaps, this is the main reason for the existence of few publicly available depression datasets. In the existing ones, because of privacy concerns, normally only labels and video features like facial landmarks, measures of facial Action Units (AUs), head pose, gaze directions, hand-crafted features, and representations of head and facial movements [43, 44, 45] are provided. To the best of our knowledge, Audio-Visual Emotion Challenge 2013 [46] and 2014 [47] (AVEC2013 and AVEC2014) datasets are the only depression datasets that currently provide raw facial video data. Another important point is the small size of the depression datasets. The long and delicate process to collect depression data may explain this fact. For instance, the Pittsburgh dataset [43] was collected from 49 patients receiving treatment for depression. Two cameras were used to capture the face and shoulders of patients and one recorded the whole body. The depression severity of each patient was evaluated in 4 clinical interviews where the interval between the interviews was 7 weeks. Furthermore, the depression datasets are usually imbalanced, having less samples for higher depression levels. To exemplify this aspect, the distribution of samples in AVEC2013 depression dataset is shown in Fig. 2.

Occurrence of similar facial behavior in depression. Depression is a heterogeneous mental health disorder which may produce ambiguities. The analysis of negative facial expressions is an example. Some studies [27, 29] report that depressed patients present reduced negative facial expressions in comparison with healthy individuals. Other studies [48, 49] report that depressed patients show more facial expressions of negative emotions. Based on both findings, it is possible to claim that negative facial expressions may be triggered with equal probability in depressed and non-depressed individuals. Therefore, depressed patients may display the same facial behavior as healthy individuals. In real-world scenarios, different facial expressions may be present

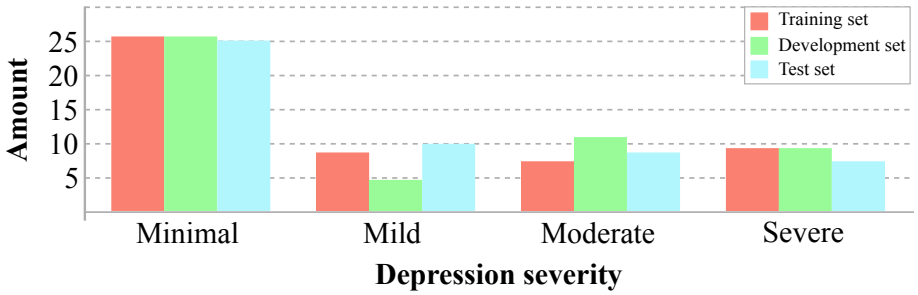


Fig. 2. The distribution of samples in AVEC2013 dataset. The number of samples for individuals with minimal depression is larger than the ones for other depression severity levels.

along distinct depression levels. For instance, a sad or neutral facial expression is exhibited in high and low levels of depression. These facts may lead to subtle differences in facial expressions along the depression levels. Moreover, the level of similarity in facial expressions may be favored by the task that a patient is performing. In AVEC2014 dataset, the individuals were recorded during a human-computer interaction process. One of the tasks performed by the individuals was to read a text, which tends to result in similar facial expressions across the depression levels.

Cultural and inter-subject variations. Cultural background may also influence facial behavior in depressive states. Across different cultures, there are different social behaviors (e.g., Latinos value more interpersonal relationships than Anglo Americans [50]) which may have an effect on emotional responses. An interesting study [50] analyzed emotional responses between depressed and non-depressed Latinas. The depressed Latinas demonstrated a similar number of facial expressions of happiness and negative emotions compared with non-depressed Latinas. This is a different facial behavior in comparison with the ones reported in other works [27, 30, 29, 48, 49] which state less positive facial expressions, and reduced or increased negative facial expressions. Regarding inter-subject diversity, there are high variations in the facial expression of an emotion displayed by individuals since the facial expressions are affected by multiple characteristics such as age, and gender [38]. It is common that the similarity between different individuals showing the same facial expression is lower than the similarity of the same individual displaying distinct facial expressions. In summary, these diversities increase the complexity for automatic understanding of the underlying factors of variation in depression data from facial information.

To overcome these challenges, this thesis proposes diverse methods to build deep learning representations for automatic depression detection from facial expressions. It is important to emphasize that the idea is not to replace clinicians or health workers. On the

contrary, the diagnostic tools are designed to assist medical professionals (psychology experts or physicians) in assessing and monitoring depressive states as well as improve accessibility for mental healthcare and reduce costs. Another potential benefit is to increase the knowledge about the correlation between depression and facial behavior (e.g., the facial behavior across diverse cultures).

1.2 Objective and research questions

The main objective of this thesis is to provide new computational models, to assist health professionals in the diagnosis and monitoring of depression, by developing novel deep learning methods that encode the correlations between facial expressions captured in videos and conditions of depression. The hypothesis is that robust and discriminative representations for automatic depression detection may be obtained from a deep architecture.

To achieve such representations, we investigate three research questions:

- Traditional screening and assessment of depression are highly dependent on clinician's subjective analysis of patient reports. Objective evaluation methods may quantify changes in facial expressions due to depressive states contributing to an accurate and reliable assessment of depression. However, exploring the correlations between facial expressions and depressive states for automatic depression detection is a non-trivial task since there are many challenges in this process, such as subtle facial expression variations along different depression levels. **In this context, the first research question (RQ1) is how to generate discriminative representations from facial expressions captured in videos for the analysis of depressive behaviors**.
- Ideally, the datasets of a healthcare application would contain multiple samples for each level of a condition, allowing a perfect description of every instance. Evidently, the elaboration of such a dataset is very difficult if not impossible. Normally, there is no large dataset for a certain medical condition. Unfortunately, this is the case of depression datasets. Specifically, the publicly available datasets are unbalanced and have a limited number of training samples. This is problematic because deep learning techniques may have high computational costs and require large amounts of data to provide good performance. When a deep learning model is trained on a small dataset, it increases the chances of overfitting, which for automatic depression detection means that the model is extracting features related to the individual instead of learning depression patterns. **Based on this discussion, the second research question (RQ2) is how to efficiently learn facial expression patterns associated to depression that allow the recognition of depression levels**.

– Depression is also associated with physical pain. A person suffering from depression may experience headache, backache, stomach ache. As pain acts as an indicator of health conditions, these physical manifestations may be considered as responses of the body to negative states of the mind. This indicates that a painful state may be informative for depression analysis. Consequently, a computational model could consider facial expressions of pain to analyze depressive states. Currently, there is no dataset that provides facial videos and annotations about pain and depression. However, it is possible to investigate the performance of methods for both applications from facial information separately, which would demonstrate the potential of the model. It is important to notice that the facial behavior of pain is different from the ones in depression. While depression is marked by the lack of expressiveness, particular facial expressions such as the ones involving closed eyes, raised cheeks, and a wrinkled nose are relevant indicators of pain [51]. **From this perspective, the last research question (RQ3) is how to encode complex patterns from facial expressions in order to obtain discriminative representations for two distinct medical applications like automatic depression detection and pain estimation**.

1.3 Contributions

This thesis contributes to objective, reliable, and contact-free solutions for the automatic assessment of depression by proposing techniques to model facial expressions in videos. The proposed methods are trained and evaluated on facial videos from depression datasets. These videos are 3D data that contain spatial and temporal information. The spatial component is related to appearance information about facial expressions. The temporal component contains the movements between frames, describing the facial dynamics. To explore facial appearance and dynamics, we introduce different deep learning techniques in order to obtain discriminative representations. In Fig. 3, we summarize our contributions, as well as their connections with the research questions and original articles.

An individual suffering from depression shows spatiotemporal alterations in his/her facial information. Thereby, the capacity of detecting and processing such changes in spatiotemporal information is important to achieve a good performance for automatic depression detection. To this end, we propose a new architecture to capture appearance and dynamic information in global and local regions of the face. With that, the detection of facial expressions related to depression may be facilitated by exploring a local facial region, and additional features may be extracted from a global facial region. In our

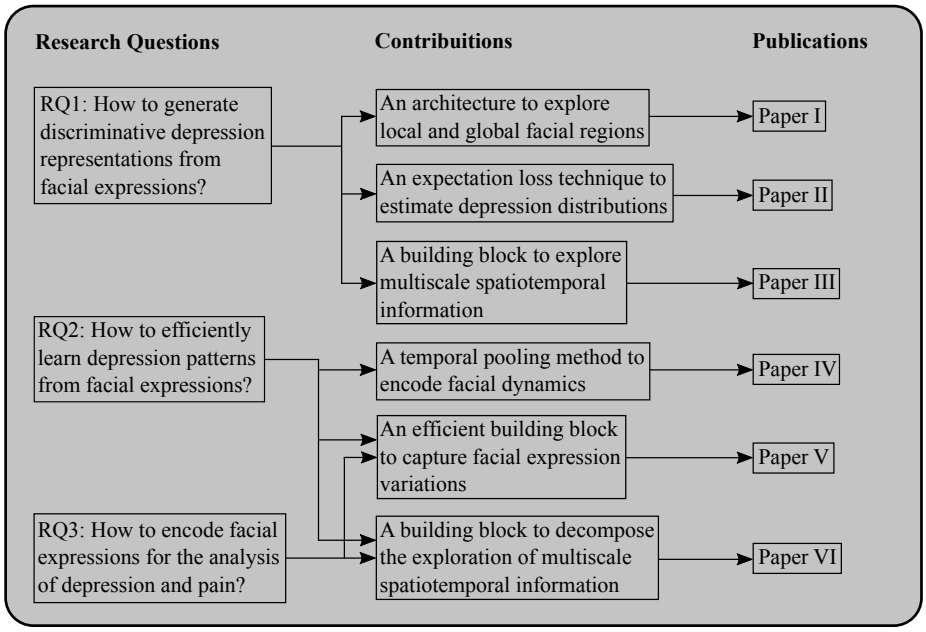


Fig. 3. The links between the research questions, and original articles, as well as their related contributions.

architecture, we employ two existent 3D CNNs to encode facial information, which are modified to improve their ability to extract spatiotemporal features.

Due to the challenges in inferring depression from facial information (e.g., subtle variations in facial expressions along different depression levels), an architecture needs to have a strong ability to model facial expressions. In this thesis, we demonstrate that the use of basic building blocks that analyze fixed spatiotemporal information limits the exploration of facial expression variations, which makes the extraction of depression features difficult. To overcome this problem, we design a new architecture using a structure that explores diverse ranges of spatiotemporal information. Experimental results indicate that this architecture has good potential to generate depression representations.

All architectures proposed in this work estimate a depression score (i.e., depression estimation). Commonly, the existing deep learning methods regard the estimation of depression scores as a regression problem by considering one single label to characterize an input sample. Such schemes do not explore the ordinal relationship between facial images and depression levels, which may result in performance degradation. In this thesis, we propose to change this procedure by using deep distribution learning.

Employing this approach, the model estimates a distribution for each input sample, and a depression score is generated from this distribution.

A common approach to generate depression representations from facial information is the use of 2D CNNs, to explore spatial information, along with an aggregation scheme, to model temporal information. One of the reasons for this fact is the availability of pre-trained models for still images. The employment of 3D CNNs is also an alternative for inference of depressive states. Both approaches have limitations to encode facial expression variations. Thus, this thesis proposes a method to encode the facial dynamics within a video into an image map. This approach favors the training process and produces an increase in performance.

This thesis further proposes efficient architectures to model facial expressions for automatic depression detection. Based on our previous study which indicated that basic building blocks with the ability to explore multiple spatiotemporal ranges contribute to the extraction of depression patterns, we investigate different strategies to design such structures. In this process, two new architectures are developed, which have the potential to generate discriminative representations for the analysis of depressive states. In the first one, the building block structure is designed to capture different facial expression variations. In the second one, the structures are designed to factorize the exploration of multiscale spatiotemporal information. One advantage of this architecture is the low computational cost, which contributes to the deployment on compact platforms with limited resources. Moreover, these two architectures further demonstrate their capacity of generating efficient representations by achieving good performance on pain estimation.

For most of the methods developed in this thesis, we present information related to the facial region more relevant for the estimation of the model, which helps to understand the decision of the models. Such information may indicate the depression patterns in facial appearance and dynamics and may be helpful for a health professional.

1.4 Summary of original articles

This thesis encompasses the development of methods to analyze facial expressions for automatic depression detection. The main focus is the generation of discriminative representations that show robustness in this challenging medical application for facial analysis. To this end, we design different deep learning techniques to explore facial information for the estimation of depression scores. The architectures developed may be divided into effective (Papers I, II and III) and efficient methods (Papers IV, V and VI).

Paper I employs Convolutional 3D (C3D) [52] to learn depression representations from facial expressions. We design an architecture that employs two C3D networks to analyze different facial regions. One of the networks explores a coarse eye region whereas the other one analyzes full-face region. A score fusion scheme combines the estimations of these networks. Interestingly, the eye region shows to hold valuable spatiotemporal information to discriminate between normal and depressive behaviors. We also integrate 3D Global Average Pooling (GAP) into C3D to improve the ability of modeling spatiotemporal information.

Paper II introduces a deep learning architecture to effectively estimate depression scores using distribution learning. Instead of building a regression model that uses a loss function to penalize the differences between the estimated and ground-truth depression, we propose a new expectation loss to estimate underlying depression distributions. The proposed method may address problems related to noisy and ambiguous labels since the model explores the relationship between facial images and depression levels. Moreover, we provide attention maps for inputs with different depression levels.

Given that frame-wise feature extraction approach has difficulty to represent dynamic information and 3D CNNs have limitations to model facial expression variations because these models explore fixed spatiotemporal information, Paper III proposes a basic building block composed of parallel 3D convolutional layers with different temporal depths and sizes of receptive field. The architecture formed by using this block has the potential to explore a wide range of spatiotemporal variations in facial expressions. Furthermore, we also present attention maps generated using this architecture.

Paper IV presents a novel temporal pooling method to capture and encode the facial dynamics inside video clips into an image map. The technique is based on binary code and the encoded image may convey information about movement and velocity in the generated texture, which favors the exploration of dynamics information by a 2D CNN model. The final architecture is a two-stream convolutional network where the temporal network explores the image map, and the appearance network explores the spatial information.

The approach in Paper III demonstrates good capability in modeling facial expressions for automatic depression detection. This result motivates an investigation of efficient methods since this approach uses parallel 3D CNNs and depression datasets are small in size. To efficiently learn depression representations, Paper V proposes an architecture that operates without 3D convolutions and explores facial expression variations at different temporal scales. The basic building block of this model is composed of a maximization block, which captures smooth transitions of facial structures, and a

difference block, which encodes sudden spatiotemporal variations. This architecture also shows promising results for pain estimation.

The architecture in Paper VI is developed to efficiently estimate depressive and painful states. Such an architecture is composed of building structures that are designed considering the facial behavior in depression and pain. In total, three building blocks are proposed, which have different abilities to explore multiscale spatiotemporal information. The employment of these blocks allows the architecture to extract a diversity of multiscale spatiotemporal features, which is important because of the different characteristics in facial expressions for pain and depression. This ability is shown to be useful for the representation learning process. In addition, Papers V and VI show attention maps produced by these architectures.

1.5 Organization of the thesis

This thesis is comprised of two parts. The first part contains five chapters that summarize the existing automatic depression detection approaches and present the original contributions. The second part is formed by the original papers.

Chapter 1 briefly introduces the background of this study and gives the objective, research questions and contributions of this work. At the end, a summary of the six original articles and the outline of the thesis are also presented.

Chapter 2 reviews depression concepts and its clinical assessment procedures. It also presents the correlations between depression and facial expressions, an overview of the existent methods for automatic depression detection, and a description of publicly available depression datasets.

Chapters 3 and 4 convey the main contributions of the thesis for estimating a subject's depression score based on facial expressions. Chapter 3 focuses on effective deep learning techniques to model facial information for this health application, including spatiotemporal representation learning methods, and a technique based on deep distribution learning. Chapter 4 focuses on the modeling of facial expressions for efficient depression representation. Different strategies are presented to learn discriminative representations, requiring reduced computational costs. This chapter also shows evaluations for pain expression recognition.

The thesis is concluded in Chapter 5 by discussing the findings of this study and giving possibilities of future works.

2 Depression: concepts, clinical diagnosis and automatic methods

In this chapter, we present a literature review about depression analysis. Firstly, we introduce the fundamental characteristics of depression and describe the standard clinical depression assessments such as self-reported questionnaires. After that, we present the evidence that depression alters the facial behavior of an individual and we review the existing methods that explore facial expressions to infer depressive states. Finally, we describe publicly available datasets for automatic depression detection from facial information.

2.1 Depression concepts

Depression is a mental health disorder defined as a negative state of mind that lasts over a long period. It may affect an individual's emotions, behavior, and physical health [53]. The World Health Organization (WHO) estimates that worldwide more than 300 million people are suffering from depression [1]. This prevalent mental disorder does not seem to have a specific cause. Some studies [54, 55] have reported alterations in terms of activity and connectivity in cortical and limbic regions (which are brain regions involved in mood regulation) of depressed individuals. Other studies [56, 57] have discovered genetic and environmental factors for depression. Stressful life episodes such as death of spouse, divorce, and unemployment have also been identified as contributors to depression [58]. Therefore, depression is considered as a result of a complex process involving biological, psychological, social and genetic factors.

Depression may also be analyzed in terms of valence and arousal. Valence informs how negative or positive is the emotional state. Arousal informs the intensity of emotion from low (calming) to high (exciting). The widely employed dimensional model [59] represents emotions as a combination of valence and arousal. In this model, depression is associated with negative valence and low arousal. Killgore [60] also observed this association in an experiment using self-report ratings of pleasure and arousal to predict self-reported depression. However, Clark *et al.* [61, 62] observed that depression could not be purely classified as negative affect. The authors identified negative and positive affect as two independent dimensions and showed evidences that a combination of relatively high negative affect and moderately low positive affect is a

better representation for depression. This definition has been commonly used in studies about depression [63, 64, 65].

2.2 Depression assessment

An accurate diagnosis of this complex disorder is important to assure appropriate treatment and to assess treatment response, thus helping to reduce the effects of the negative state of mind. Clinical depression evaluations usually rely on self-report questionnaires or instruments administered by a clinician. Most commonly used methods are Patient Health Questionnaire (PHQ-9) [24], Beck Depression Inventory II (BDI-II) [66] and Hamilton Rating Scale for Depression (HAMD-17) [25].

The PHQ-9 is a nine-item self-report inventory that has been extensively employed in primary care [67]. These items (see Fig. 4) are based on the diagnostic criteria of DSM-5 for depression. For each item, a rating of 0, 1, 2 and 3 represent "Not at all", "Several days", "More than half the days", and "Nearly every day", respectively. A patient responds to the questions considering symptoms within the last 2 weeks. The total score of PHQ-9 for the nine items ranges from 0 to 27. With that, four depression severity levels can be estimated:

– Minimal depression (score range of $0-9$)
– Mild depression (score range of $10-14$)
– Moderately severe depression (score range of $15-19$)
– Severe depression (score range of $20-27$)

In general, PHQ-9 is an instrument that has few items and is simple to assign scores. The instrument is freely available and a patient usually administers it, but a physician may also use it.

The BDI-II is a self-report questionnaire for assessing the existence and severity of depressive symptoms. The BDI-II contains 21 items where every item has four statements describing distinct degrees of symptom severity experienced over the past two weeks. This inventory is also based on DSM-5 criteria. When an individual is using BDI-II, a value ranging from 0 to 3 is selected in each question, which is linked to the statement that best describes his feeling. The value 0 denotes an absence of a symptom whereas the values 1, 2, and 3 indicate mild, moderate and severe symptoms, respectively. The total score is defined by adding the values of the 21 items, which may be a value ranging from 0 to 63. According to the BDI-II scores, four depression severity levels may be estimated:

– Minimal depression (score range of $0-13$)

Question	Score
Little interest or pleasure in doing things?	0 1 2 3
Feeling down, depressed, or hopeless?	0 1 2 3
Trouble falling or staying asleep, or sleeping too much?	0 1 2 3
Feeling tired or having little energy?	0 1 2 3
Poor appetite or overeating?	0 1 2 3
Feeling bad about yourself - or that you are a failure or have let yourself or your family down?	0 1 2 3
Trouble concentrating on things, such as reading the newspaper or watching television?	0 1 2 3
Moving or speaking so slowly that other people could have noticed? Or the opposite - being so fidgety or restless that you have been moving around a lot more than usual?	0 1 2 3
Thoughts that you would be better off dead, or of hurting yourself in some way?	0 1 2 3

Fig. 4. Questions of the Patient Health Questionnaire (PHQ-9).

– Mild depression (score range of $14 - 19$)
– Moderate depression (score range of $20 - 28$)
– Severe depression (score range of $29 - 63$)

The BDI-II has been applied for research analysis and clinical experience and is normally employed in psychiatric settings [68]. A disadvantage of BDI-II is that this instrument is not free.

The HAMD-17 is an observer-administered instrument comprising 17 items. A specialist selects for each item an option that best describes the patient. The items related to depressed mood, feelings of guilt, suicide, work and activities, retardation, agitation, anxiety psychic, anxiety somatic, hypochondriasis, and loss of weight are evaluated using a number from 0 to 4. The items that cover insomnia: early in the night, insomnia: middle of the night, insomnia: early hours of the morning, somatic symptoms gastro-intestinal, general somatic symptoms, genital symptoms, and insight are evaluated using a value ranging from 0 to 2. A final score is obtained by adding the values of the 17 items, which means a score ranging from 0 to 54. Based on this score, four depression severity levels may be determined:

– Absence of depression (score range of $0 - 7$)
– Mild depression (score range of $8 - 16$)
– Moderate depression (score range of $17 - 23$)
– Severe depression (score range of $24 - 54$)

Studies [69, 70] consider a positive response to depression treatment when there is a reduction of 50% or more in terms of HAMD-17 score and define remission as a HAMD-17 score of 7 or lower. Moreover, a specialist may complete the questionnaire in 20 to 30 minutes [25].

In summary, the diagnosis of depression depends on clinician's perception of verbal reports from individuals or questionnaires completed by the patients. However, studies have reported a high number of misdiagnoses of depression by clinicians [22, 23, 71], with potentially serious consequences for the patients [23]. To aid health professionals, automatic objective methods may be developed to analyze behaviors related to depression. In particular, systems that explore facial behaviors which may use contact-free sensors to provide unobtrusive solutions for the diagnosis and monitoring of patients.

2.3 Analyzing depression from facial expressions

The human face is a natural source of information. It may indicate age, personality, emotions, attractiveness, etc. Face may also convey information about health status

because behavioral responses to certain clinical conditions modify the expressions of the face [72]. Such alterations may be an important information for clinicians and/or automatic systems.

2.3.1 Facial expressions

The face is composed of multiple muscles which contribute to different processes like speaking. The movements of facial muscles produce a facial expression and the diverse arrangements of these movements form different facial expressions. These facial changes are usually associated with reactions to internal states of an individual (e.g., emotions or intentions) [73]. Given that these facial motions are very informative, various studies have been conducted on facial expression analysis. They may broadly be divided into universal facial expressions, and Facial Action Coding System (FACS). The first one refers to the studies based on the work of Ekman and Friesen [74, 75] who proposed that there is an prototypical facial expression, which is universally recognized, corresponding to a basic emotion. Anger, disgust, fear, happiness, sadness, and surprise are the perceived basic emotions. However, recent studies show that these emotion-specified expressions are not universal but culture-specific [76], and in real world scenarios these prototypical facial expressions rarely occur [73].

FACS [74] is a different approach that was designed to identify motion in facial muscles. The system represents facial expressions using Action Units (AUs) which are a visually distinguishable action of an individual muscle or group of muscles. To generate this representation, the FACS coders examine the facial expressions and determine the beginning (onset), intensity, and ending (offset) of the AUs. In total, the FACS employs 44 AUs to describe facial expressions. The activity of a specific set of facial muscles are described in 30 AUs where 12 are related to upper face region and 18 belong to lower face region. Table 1 details these AUs showing the FACS names and muscular basis. The remaining AUs are classified as miscellaneous actions.

2.3.2 Correlations between depression and facial expressions

Psychiatrists, psychologists, physicians, and clinical researchers as well as computer scientists have identified variations in facial expression associated with depression. Some authors recognized facial behaviors in depressive conditions by analyzing psychomotor retardation [8, 13, 77, 78], which is one of the major features of depression. Their works found that the manifestations of this symptom include slowed speech rate, flat facial expression, poor eye contact, etc. In one of these works, Widlöcher [77] explained that

Table 1. Description of thirty action units.

AU	FACS name	Muscular Basis
1,2	Inner, Outer brow raiser	Frontalis
4	Brow lower	Depressor glabellae, depressor supercilii, corrugator
5	Upper lid raiser	Levator palpebrae superioris
6	Cheek raiser	Orbicularis oculi
7	Lid tightener	Orbicularis oculi
9	Nose wrinkler	Levator labii superioris, alaquae nasi
10	Upper lip raiser	Levator labii superioris, caput infraorbitalis
11	Nasolabial deepener	Zygomaticus minor
12	Lip corner puller	Zygomaticus major
13	Sharp lip puller	Levator anguli oris
14	Dimpler	Buccinator
15	Lip corner depressor	Depressor anguli oris
16	Lower lip depressor	Depressor labii inferioris
17	Chin raiser	Mentalis
18	Lip pucker	Incisivii labii superioris and inferioris
20	Lip stretcher	Risorius with platysma
22	Lip funneler	Orbicularis oris
23	Lip tightener	Orbicularis oris
24	Lip pressor	Orbicularis oris
25	Lips part	Orbicularis oris
26	Jaw drop	Masseter; relaxed temporalis and internal pterygoid
27	Mouth stretch	Pterygoids, digastric
28	Lip suck	Orbicularis oris
41	Lid droop	Relaxation of levator palpebrae superioris
42	Slit	Orbicularis oculi
43	Eyes closed	Relaxation of levator palpebrae superioris; orbicularis oculi, pars palpebralis
44	Squint	Orbicularis oculi, pars palpebralis
45	Blink	Relaxation of levator palpebrae and contraction of orbicularis oculi, pars palpebralis
46	Wink	Relaxation of levator palpebrae superioris; orbicularis oculi, pars palpebralis

the depressed patients present a scarcity of head movements, and they look fixedly at the ground. He also described the facial expressions of the patients as motionless and blank. Other studies observed that during clinical interviews patients with depression have an increase in frowning [34, 35, 37]. In this case, Stratou *et al.* [35] observed a different behavior between men and women. Employing AU 4 intensity as a measure of frown, the authors showed results indicating that depressed men have more frowns than non-depressed men, but depressed women show opposite behavior. Smiles have also been observed for the analysis of depression [31, 32, 37, 79]. Fairbanks *et al.* [37] reported a reduction in the number of smiles in psychiatric patients compared to healthy individuals. Scherer *et al.* [31, 32] found that depressed patients display smiles with a shorter duration and less intensity. Girard *et al.* [79] demonstrated that patients with severe depression levels smile less often and these smiles were frequently associated with facial actions related to contempt (AU 12 and AU 14). Decreased mouth movements have also been observed in depressive states [33, 9].

A common approach to examine facial expressions related to depression is to induce either positive or negative emotions in the patients by showing short movie sequences. Using this approach, Renneberg *et al.* [27] compared the behavior of 27 women suffering from depression and 30 healthy female individuals. When induced to positive mood, depressed patients showed notably less emotionally positive facial expressions than the healthy group. When induced to negative mood, depressed patients displayed significantly less negative facial expressions than the healthy group. Based on that, the authors stated that depressed patients show diminished facial emotional expressiveness. Similarly, Berenbaum *et al.* [30] reported depressed patients showing less facial responses to positive and negative stimuli than healthy individuals. Gaebel *et al.* [29] analyzed facial reactions during an emotion inducing interview. Their results also show evidences of a reduced facial expressiveness in depression. On the other hand, Sloan *et al.* [28] reported reduced positive facial expressions in depression, but no significant difference between health and depressed individuals in terms of negative facial expressions. Other works also indicated different results about emotionally negative facial expressions. Sloan *et al.* [48] reported that depressed patients show more negative facial expressions than health individuals. Brozgold *et al.* [49] found that depressed patient display more negative and less positive facial expression than the healthy group. Tsai *et al.* [50] reported different results. These authors observed emotional responses of 12 depressed and 10 non-depressed Latina women. Their results showed no significant difference between the two groups in facial expressions of happiness or negative emotion, but depressed Latinas displayed fewer social smiles

during positive stimuli. Despite the small sample size, this work indicates the importance of investigating depression in a cross-cultural context.

2.4 Automatic depression detection from facial information

As described in Section 2.2, the diagnosis of depression is subjective in nature. It relies on the opinion of either the clinician performing the analysis or the patient himself. This fact and the clinical evidence indicating correlations between facial expressions and depression (see Section 2.3) have motivated the development of automatic depression detection systems. Such medical assistive tool analyzes facial variations for objective recognition of expressions related to depression which may contribute to the improvement of clinical assessment and monitoring. This task is considered to be a challenging recognition problem since the facial expression variations may be small along different depression levels. Two other important aspects are the limited size of depression datasets and a possible change in facial behavior across different cultures. Moreover, the analysis of facial expressions of individuals has inherent challenges such as large intra-class variations and large inter-class similarities [38]. Consequently, a computational model must produce discriminative and robust representations to accurately infer depressive states from facial information. In this thesis, diverse strategies are developed to analyze facial expressions with the aim at generating such representations for automatic depression detection.

Diverse methods have been proposed to explore facial expressions for estimation of depressive states. The fundamental stage in these schemes is the feature extraction since it detects and encodes patterns related to depression in facial expressions. Based on the feature extraction method employed, we divided the existent methods into four groups: histogram based techniques, motion based techniques, deep learning based techniques, and facial action units based techniques.

2.4.1 Histogram based techniques

This class of methods is used to describe local features of facial expressions. One of the most popular descriptors is Local Binary Pattern (LBP) [80]. This algorithm produces a value for each element in the input image by thresholding a neighborhood using the value of the central pixel, which generates a binary pattern. The LBP descriptor employs a histogram of these values to represent the facial image. Tadalagi *et al.* [81] proposed a system to analyze facial expressions for depression severity recognition using LBP descriptor. The system applied LBP to extract features from static images and

performed classification using Support Vector Machine (SVM) method. A disadvantage of the LBP descriptor is that the method only explores spatial information. Given this limitation, different variants have been developed such as Local Binary Patterns on Three Orthogonal Planes (LBP-TOP) [82], and Local Gabor Binary Patterns from Three Orthogonal Planes (LGBP-TOP) [83] which may be applied for spatiotemporal modeling of facial expressions. LBP-TOP employs LBP in XY, XT, and YT planes, where X, and Y are spatial coordinates and T is the temporal dimension, then it concatenates the statistics of three planes to form a histogram. LGBP-TOP may be seen as a combination of Gabor wavelet filtering and LBP-TOP. Gupta *et al.* [84] used LBP-TOP descriptor to explore dynamic information, then employed Support Vector Regressor (SVR) method for depression estimation. Dhall *et al.* [85] extracted LBP-TOP features and encoded these features by using Fisher Vector (FV) method. Valstar *et al.* [47] proposed a system based on LGBP-TOP features which was used as baseline in the Audio-Visual Emotion recognition Challenge 2014 (AVEC 2014). In [86], the authors employed Principal Component Analysis (PCA) to reduce the dimensionality of LGBP-TOP features, then used Relevance Vector Machine (RVM) method for depression estimation.

Another common visual descriptor is Local Phase Quantization (LPQ) [87]. This histogram-based feature extractor employs decorrelation and quantization schemes on Fourier transform phase in local rectangular regions. To estimate depression levels, Kächele *et al.* [88] used LPQ descriptor for feature extraction followed by SVR. Valstar *et al.* [46] defined LPQ features as appearance baseline features in AVEC 2013. In order to capture spatiotemporal information, a variant of this descriptor called Local Phase Quantization from Three Orthogonal Planes (LPQ-TOP) [89] was developed using a technique similar to LBP-TOP. Wen *et al.* [90] demonstrated that LPQ-TOP features are more discriminative than LPQ features for depression estimation. Histogram of Oriented Gradients (HOG) which concatenates histograms of gradient orientations from different areas of the image has also been investigated for automatic depression detection [91]. Moreover, it is possible to improve the representations by combining different features. For instance, Kaya *et al.* [92] combined LGBP-TOP and LPQ features using Canonical Correlation Analysis (CCA) method, generating an improvement in performance.

The aforementioned methods have presented limited capabilities in producing effective representations for the analysis of depressive states. Indeed, it is difficult to manually design a technique that detects and extracts depression patterns from facial expressions.

2.4.2 Motion based techniques

Temporal information is essential for the automatic analysis of depression. An approach that considers only the spatial information has to deal with high level of ambiguity, e.g., negative facial expressions may be displayed in an image by depressed and non-depressed individuals. A motion based technique captures the dynamics of facial expressions by computing different measures of motions in a sequence of frames. Pérez *et al.* [93] investigated a motion based technique for depression estimation from face videos. The method divides videos into segments where face and eyes are detected. To measure the movements of the face and eyes along the segment, the minimal, maximal, and average values of coordinates of the face and eyes are calculated. The distance between the initial and final positions of the face and eyes are also computed as well as the average velocity of face and eyes.

Other more sophisticated methods have been proposed to encode the dynamic information of videos. Meng *et al.* [94] investigated Motion History Histogram (MHH) to extract dynamic features from face videos for depression estimation. The algorithm registers different levels of motion in a fixed number of grayscale images, where each image contains information about a specific level of motion. The authors further improved the MHH features by using either LBP descriptor or Edge Orientation Histogram (EOH) [95]. The final representation of the method was the combination of histograms from LBP and EOH. For depression detection, Maridaki *et al.* [96] explored Motion History Image (MHI) [97] and Gabor Motion History Image (GMHI) [10] methods. MHI algorithm encodes motion patterns into a grayscale image by associating pixel intensities with temporal history of movement. GMHI algorithm employs the same procedure as MHI, but the video frames are first processed by a bank of Gabor filters. The authors investigated several combinations involving MHI, GMHI, LBP, LPQ and HOG, where the combination of MHI and HOG, and the scheme that computed statistics directly from either MHI or GMHI features such as mean, standard deviation, and histograms provided better representations. Another method for motion estimation is optical flow algorithm which registers velocity of pixels along consecutive frames into an image. Zhu *et al.* [98] generated flow images by considering movements between two frames, with interval of 10 frames. The authors employed a deep learning technique to explore the flow images for depression estimation.

Despite that motion-based techniques have the capacity to explore facial dynamics, problems related to motion self-occlusions, and illumination variations may cause difficulties to represent the facial information. Moreover, considering the variation between two consecutive frames may generate similar representations along different

levels of depression, which means that this approach may reduce the discriminative ability of the method.

2.4.3 Deep learning based techniques

Deep learning models have also been employed for automatic depression detection. This approach has provided state-of-the-art performance in many visual recognition applications [99, 100, 101], including automatic facial analysis [39, 41, 40]. In general, deep models use Convolutional Neural Networks (CNNs) to learn representations from videos. These methods may generate depression representations by either using frame-wise feature extraction or exploring directly the spatiotemporal information. The frame-wise approach employs 2D CNNs to explore spatial correlations along with an aggregation scheme or a recurrent technique to capture temporal dependencies. Jan *et al.* [102] employed a deep network with 16 convolutional layers and 3 fully connected layers to extract spatial features, followed by the Feature Dynamic History Histogram (FDHH) algorithm to model variations of the static features. Kang *et al.* [103] used a pre-trained 2D CNN to extract facial features, then a Deep Transformation Learning (DTL) scheme was employed to project these features into a new subspace for estimation of depression scores. Zhou *et al.* [104] employed ResNet-50 architecture to explore appearance information. To encode the temporal dependencies, the authors proposed a memory attention mechanism which consists of two cascaded attention blocks that learn weights to combine the facial features. The last stage of the method is a fully connected layer that outputs the depression scores. In [105], the authors developed a system that uses four ResNet-50 to explore different facial regions in images, aiming to recognize the informative facial regions in depressed individuals. The system computed the depression score by averaging the estimations of each frame belonging to the video under analysis.

To directly leverage the spatiotemporal information across frames, the systems may employ 3D CNNs. Al Jazaery *et al.* [106] used a 3D CNN method with eight $3 \times 3 \times 3$ convolutions and two fully connected layers to capture spatiotemporal variations. Then, a Recurrent Neural Network (RNN) was used to provide estimations. The authors designed the final architecture with two 3D CNNs in order to explore the facial region and a region that includes the head area. In this way, the architecture may analyze facial expression variations and head movements. Another possibility to directly explore the spatiotemporal information is the use of two-stream networks. In this strategy, the appearance information is explored by one network whereas the other one is responsible for capturing facial dynamics. In both cases, the networks may be 2D CNNs that

were pre-trained on large datasets. Zhu *et al.* [98] employed two-stream networks for depression estimation where the temporal network explored optical flow images. The authors used a fusion scheme to produce a depression score.

In summary, most of the methods that use deep learning techniques are based on frame-wise feature extraction. However, the intrinsic spatiotemporal correlations may deteriorate using this type of technique [39]. Few methods have been proposed to explore directly the information between frames to infer depressive states which motivates the development of new strategies using this approach.

2.4.4 Facial action units based techniques

This approach explores the characteristics of facial action units to infer depressive states. Cohn *et al.* [107] proposed a system that explored 17 manually annotated facial AUs obtained during clinical interviews and used SVM as a classifier for depression detection. For each AU, the method computed average duration, the proportion of occurrence in relation to total time of the interview, the ratio between onset phase and total duration, and the ratio between onset phase and offset phase. The authors reported that using all 17 facial AUs the system achieved 79% accuracy, but, employing only AU 14, the accuracy was even higher. This study also provided a comparison of this method with the model based on Active Appearance Model (AAM) features, where the results showed competitive performance. Similarly, Girard *et al.* [79] investigated the use of manual and automatic coding of facial AUs to analyze facial behavior in depression, where their results demonstrated consistency between these two approaches. The findings of these two studies were important to indicate that automatic systems have the potential to identify depression patterns in facial expressions.

In order not to rely on annotators to measure AUs, the systems normally employ an existent toolbox to detect intensities of facial AUs in video frames. Williamson *et al.* [108] proposed a multimodal approach for depression estimation where the facial information was explored by using the Computer Expression Recognition Toolbox (CERT) [109] to detect 20 AUs. Song *et al.* [110] employed OpenFace 2.0 toolkit to detect intensities of 17 facial AUs. Specifically, the method used this tool to detect AUs, head pose, and gaze directions. These signals were transformed to the frequency domain, then the spectral representations were explored by 1D CNNs to estimate a depression score. The findings of this work suggested that AU 4, AU 12, AU 15, and AU 17 are informative for depression estimation.

The analysis of depression using manually annotated AUs is difficult because facial AU annotation is a time-consuming, and expensive process. On the other hand, the

problem of developing a method for automatic depression detection using available tools to detect the presence and intensity of AUs is that the performance of this method is affected by the performance of the tool.

2.5 Depression datasets

In the development of methods for automatic depression detection, data is an essential part because it contains the depression patterns. The scientific community have created some depression datasets in order to contribute to the progress of automatic analysis of depressive behaviors. In general, making a depression dataset is a long and delicate process. The first step is the recruitment of individuals in a hospital, mental health clinic or through websites. In the case of individuals under treatment for depression, normally the creators of the dataset select participants that satisfy DSM-5 criteria [43, 79, 107]. To register the depressive behavior, the participants are recorded in a Human-Computer Interaction (HCI) task or in an interview conducted by clinicians or virtual human interviewer. The participants may be recorded more than once, with a period of weeks between the interviews. From the technical perspective, important configurations have to be carefully defined such as illumination conditions, number of cameras, positions of the cameras, resolution of the cameras, and regions of interest.

A clinician may measure the depression severity or responses to treatment of a participant by using HAMD-17 instrument. Another option is the employment of self-report questionnaires like BDI-II and PHQ-9. Consequently, the ground truth of the datasets may be based on different depression scores (e.g., BDI-II score). Furthermore, there exists a reasonable number of depression datasets, but few of them are publicly available. In the following, a brief presentation of the publicly available depression datasets is provided.

2.5.1 AVEC 2013 dataset

The Audio-Visual Emotion Challenge and Workshop held in 2013 [46] and 2014 [47] (AVEC 2013 and AVEC 2014) has contributed significantly to the development of automatic systems for depression analysis. In both events, one of the tasks of the participants was to estimate self-reported depression scores in videos. The datasets employed in the competitions are named AVEC 2013 and AVEC 2014 depression datasets and are publicly available for research purposes.

We firstly describe AVEC 2013 dataset [46]. This depression dataset is a subset of the Audio-Visual Depressive language Corpus (AViD-Corpus). It consists of individuals

performing HCI tasks while being recorded by a webcam and a microphone. Specifically, Power Point was used to guide the individuals to carry out a series of activities. Some of these activities were sustained vowel phonation, sustained loud vowel phonation, sustained smiling vowel phonation, speaking out loud while solving a task, counting from 1 to 10, telling a story from the individual's own past. In total, there are 150 videos where each one of them contains only one individual. The average length of the videos is 25 minutes where the shortest video reaches 20 minutes in duration, and the longest one lasts 50 minutes. The frame rate of the videos is 30 frames per second, and the frame resolution is 640×480 pixels.

The videos are allocated into three different partitions: training, development and test sets. Every set comprises 50 videos where each one of them has a label corresponding the depression score of an individual. Given that the BDI-II was used as assessment instrument for this dataset, a BDI-II score is defined as the ground truth in each video. Consequently, the depression scores range from 0 to 63 where the average depression score is 15.1 and 14.8 for training and development sets, respectively. In Table 2, we report the distribution of BDI-II scores in each set of AVEC 2013 depression dataset.

The organizers of the event made the videos (audio and images), used in the competition, publicly available. During the competition, the organizers provided LPQ features as video baseline features to the participants. These features are also provided in the AVEC 2013 dataset.

2.5.2 AVEC 2014 dataset

The AVEC 2014 depression dataset [47] is also a subset of the AViD-Corpus. The individuals were recorded during HCI tasks by using a webcam and a microphone. For this dataset, the individuals performed two different activities: Northwind, and Freeform tasks. In the Northwind task, the individuals were asked to read an excerpt of the fable "The North Wind and the Sun". In the Freeform task, the individuals discussed a bad or good childhood memory as well as responded different questions like "What was your best gift, and why?". This process resulted in 300 videos where each sample registers the behavior of one individual. In comparison with AVEC 2013 dataset, the videos in AVEC 2014 dataset are shorter. The length of the videos ranges between 6 and 248 seconds. The videos have frame rate of 30 frames per second, and the frame resolution is 640×480 pixels.

For each task, the videos are allocated into training, development, and test sets. Every set contains 50 videos where each sample is labeled with a depression score, denoting the depression severity of an individual. Since the assessment instrument used

Table 2. Distribution of BDI-II scores in the sets of AVEC 2013 and AVEC 2014 depression datasets.

Sets	Severity	Score range	Number of samples	
			AVEC 2013	AVEC 2014
Training	minimal	0 – 13	26	52
	mild	14 – 19	8	16
	moderate	20 – 28	7	12
	severe	29 – 63	9	20
Development	minimal	0 – 13	26	54
	mild	14 – 19	4	8
	moderate	20 – 28	11	20
	severe	29 – 63	9	18
Test	minimal	0 – 13	25	50
	mild	14 – 19	10	20
	moderate	20 – 28	8	14
	severe	29 – 63	7	16

for this dataset was BDI-II, the ground truth of each video is the BDI-II score of the individual in analysis. Moreover, the average depression score of the individuals in the training set is 15.0, whereas the average for development set is 15.6. The distribution of BDI-II scores over the sets of AVEC 2014 depression dataset is summarized in Table 2.

The organizers of this event also made the videos, employed by the participants of the competition, publicly available. The organizers defined LGBP-TOP as video baseline features and also provided these features in the AVEC 2014 dataset. Until the date of writing of this thesis, AVEC 2013 and AVEC 2014 datasets are the only depression datasets that provide raw video information.

2.5.3 DAIC-WOZ dataset

The Distress Analysis Interview Corpus – Wizard of Oz (DAIC-WOZ) depression dataset [45] is a subset of DAIC corpus, which has a collection of clinical interviews that were designed to support the diagnosis of depression as well as other psychological distress conditions. The goal of the interviews was to develop a computer agent with the ability to interview individuals and recognize verbal and non-verbal behaviors related to

mental disorders. The dataset contains information that was collected during Wizard of Oz interviews, which means that a virtual human interviewer called Ellie conducted the interviews.

The dataset is formed by 189 sessions of interactions with range from 7 to 33 minutes, where the average length of the sessions is 16 minutes. These sessions are allocated into training, development, and test sets. The training set contains 107 samples, while the development, and test sets have 35 and 47 samples, respectively. The PHQ-8 is the assessment instrument used to measure the depression severity. This instrument is similar to PHQ-9 except for item 9 and score range of severe depression. For each sample, DAIC-WOZ dataset provides a PHQ-8 score, and a binary value indicating if the individual is depressed (PHQ-8 score ≥ 10) or not depressed. The distribution of binary labels across the sets of DAIC-WOZ dataset is reported in Table 3.

Instead of raw video recordings, DAIC-WOZ dataset provides facial features that were obtained using OpenFace tool kit [111]. The available features are 68 2D and 3D facial landmarks, HOG features, estimation of gaze directions for both eyes, orientation of the head, 3D position of the head, and facial activations (AU 1, AU 2, AU 4, AU 5, AU 6, AU 9, AU 10, AU 12, AU 14, AU 15, AU 17, AU 20, AU 25, AU 26). Moreover, this dataset was employed in the AVEC sub-challenge held in 2016 [44] and 2017 [112].

Table 3. Class-wise distribution of DAIC-WOZ depression dataset.

Sets	Number of samples		
	Depressed	Non-depressed	Total
Training	30	77	107
Development	13	22	35
Test	14	33	47

2.5.4 Pittsburgh dataset

The Pittsburgh depression dataset [43] contains samples obtained from 49 individuals that were recruited during a clinical trial for treatment of depression. These individuals were interviewed by a group of 10 specialists. To record the clinical interviews, the initial setup employed for this dataset was four analogue cameras synchronized by hardware, where two cameras registered the face and shoulders of an individual, one camera registered the full body of an individual, and one camera registered the interviewer [107]. However, the final setup employed for video recordings disregards one of the cameras used to capture the face and shoulders of an individual [43]. The recordings were

converted into a digital format of 640×480 pixels at rate of approximately 30 frames per second.

During the treatment, the depression severity level of each individual was assessed a total of four times at 1, 7, 13, and 21 weeks. The assessments were performed using HAMD-17 instrument, which is administrated by a clinician. For this reason, the samples of Pittsburgh dataset are labeled based on HAMD score. Moreover, the study identified the health status of the individuals by mapping the HAMD scores into three different severity levels or classes: remission (HAMD scores of 7 or lower), mild depression (HAMD scores between 8 and 14), and moderate to severe depression (HAMD scores of 15 or higher). In this way, the dataset may be employed for depression severity recognition.

In total, the number of samples on Pittsburgh dataset is 130. The distribution of these samples per class is reported in Table 4. The dataset does not provide the video data recorded during the clinical interviews. Alternatively, it includes visual features that were extracted from raw video data. Specifically, these features are normalized 2D coordinates of 49 facial landmark points, 3 degrees of freedom of head pose (i.e., pitch, yaw, and roll), and deep features extracted from head and facial movement using stacked denoising autoencoder method.

Table 4. Class-wise distribution of Pittsburgh depression dataset.

Severity	Score range	Number of samples
Remission	$0-7$	37
Mild	$8-14$	35
Moderate to Severe	$15-35$	58

2.6 Summary

Depression is a common mental health disorder characterized by a persistent negative state of mind. A person suffering from depression may have diverse emotional and physical problems that may lead to suicidal behavior. An accurate evaluation of depression and its severity is essential to start proper treatment and to assess treatment response so that the patient may recover from this disorder. A challenge in assessing depression is that this process relies on clinicians' opinions and reports from patients. To improve this procedure, automatic diagnosis systems may be employed to assist the diagnosis and monitoring of patients. These systems may explore facial

expressions, which contain rich information about depression and allow the development of unobtrusive solutions.

Previous studies have shown evidence that there are alterations in facial expressions due to depression. An objective assessment method may automatically analyze these changes to infer depressive states. For example, slowed speech rate, smiles with a shorter duration and less intensity, and a small number of mouth movements are some cues related to depression. Consequently, a system may quantify the variations in the mouth region to detect depression and its severity. Considering also other facial cues, such as motionless facial expressions, diminished facial emotional expressiveness, reduced positive facial expressions, etc, it is possible to understand that the appearance and dynamics of facial expressions convey essential information related to depression.

Different approaches have been proposed to encode the facial information for depression detection, depression severity recognition, and depression estimation. Various methods employ histogram based algorithms to extract features from static images or videos. Motion and facial action units based algorithms have also been used to represent depressive states. The features generated by these methods are explored by conventional classifiers or regression schemes (e.g., SVM, and SVR) to generate an output. Recently, deep learning based methods have demonstrated more potential to extract discriminant features for automatic depression detection from facial expressions. Usually, the methods employ a sequence of convolutional layers, as the feature extractor, and fully connected layers to generate a depression score or to determine a depression level. Most of these methods use 2D CNNs to explore appearance information and some aggregation technique to explore facial features that are extracted from videos. Such approach has limited ability to encode important dynamic information. Few methods explore directly spatiotemporal variations in facial information. Therefore, an investigation of different strategies to model facial expression variations is important in order to understand how to effectively generate depression representations from videos.

Deep learning models require a massive amount of training data to identify patterns in data. When it comes to automatic analysis of depression, these models explore the data to learn correlations between facial expressions and depression, so the dataset used in this process plays an important role. However, the publicly available depression datasets are unbalanced and small in size. When a deep model is trained on such datasets, the chance of overfitting increases considerably. Therefore, the development of efficient deep architectures, which require the optimization of reduced trainable parameters, may benefit the generation of depression representations.

3 Learning depression representations from facial expressions

The recent advances in deep learning and computer vision techniques have contributed to the development of objective methods for video-based automatic depression detection. Many existing methods address the recognition of depressive states by analyzing facial expressions. Such an approach tries to find facial patterns that allow the identification of depressed and non-depressed individuals, as well as the distinction of different depression levels. Consequently, the development of methods that have the potential to produce powerful representations from visual data has been the main point of focus for research in depression analysis. In this chapter, we summarize our contributions in Papers I, III, and II which present different techniques to effectively generate depression representations from facial expressions.

3.1 Introduction

Facial expressions are a valuable non-verbal source of information for video-based depression analysis. An accurate modeling of this facial information may contribute to the advancement of assistive medical tools for the recognition of depression behaviors. The modeling of facial expressions involves exploring correlations of this information in order to learn discriminative representations. One possible approach to such modeling is to analyze dependencies from the appearance of facial expressions. However, such an approach hinders the modeling since depressed individuals may display the same facial expression of emotion as healthy individuals. For this reason, the analysis of temporal information is crucial to model facial expressions in depression conditions. For instance, depressed and non-depressed individuals show positive facial expressions, but depressed individuals display less expressions related to positive emotions [27, 29, 30]. Thus, the observation over time helps to understand the differences in facial behavior between depressed and non-depressed individuals. Therefore, the modeling of facial expressions must take into account appearance and dynamic information to generate effective depression representations.

The modeling of facial expressions using spatial and temporal information for the estimation of depression scores has been approached through conventional and deep learning methods. The solutions based on conventional methods use hand-crafted features to represent depressive states [47, 84, 85, 86, 89, 90, 92, 94, 96]. This

approach has demonstrated limited ability to explore spatiotemporal dependencies, which deteriorates the estimation performance. Deep learning methods have shown stronger ability to explore spatiotemporal information for depression analysis [98, 102, 103, 104, 105, 106]. Although these deep methods achieve a good level of performance, the similarity of facial expressions along different depression levels remains a key challenge. Indeed, as discussed in Chapter 2, an individual who is suffering from depression displays positive and negative facial expressions of emotion, which may lead to samples with small differences in facial expressions, but with distinct depression severity levels.

In this chapter, we introduce different deep learning techniques to explore spatiotemporal information in order to improve the modeling of facial expressions for depression estimation. We also propose to investigate the robustness of depression distributions to estimate a depression score for an individual in video. Moreover, we provide experimental results to elucidate the contributions of the proposed methods and a comparison with the state-of-the-art methods for depression estimation to demonstrate the capabilities of our methods.

3.2 A global-local convolutional 3D

The extraction of depression patterns from the appearance and dynamics of facial expressions is a challenging process since there are intrinsic complexities related to depression, and difficulties regarding automatic facial analysis such as variability of different subjects, capture conditions, large intra-class variations, and large inter-class similarities. To address these challenges, deep learning techniques are usually designed using 2D CNNs to explore spatial information, along with an aggregation method to capture temporal information [102, 103, 104, 105]. Consequently, these techniques explore spatial and temporal information separately. However, this approach has shown limited capacity in encoding important dynamic information [39].

The above discussion motivates an investigation of different approaches to achieve robust automatic depression assessment from facial expressions. To this end, we design a new deep architecture to enhance the extraction of depression features considering two aspects: 1) encoding spatial and temporal information simultaneously is a better approach to generating depression representations, and 2) a local facial region may convey relevant information to the identification of emotional states [113]. Based on these two points, the proposed architecture is designed using 3D CNNs to leverage spatiotemporal relationships of local and global facial regions. More precisely, we employ two Convolutional 3D (C3D) networks to model facial expression variations of

these facial regions. The rationale behind the use of C3D to model variations from a local region is that it increases the focus on a region that may show essential information for depression analysis.

3.2.1 C3D network

The C3D network has been successfully employed in action, scene, and object recognition [52]. The model have gained popularity due to the availability of C3D networks that were pre-trained on large-scale datasets and its good capacity to encode spatiotemporal information. In order to use a pre-trained C3D network to estimate the depression score of an individual in video, the network has to be fine-tuned on a depression dataset through the process called transfer learning. In this way, the knowledge acquired from another domain (e.g., action recognition) may be used to improve the learning process for depression estimation. In our approach, we employ C3D networks that are pre-trained on Sports-1M [114] and UCF101 [115] datasets, then the networks are fine-tuned on AVEC2013 and AVEC2014 depression datasets.

C3D is composed of 8 convolutional layers, and 5 pooling layers in the feature extraction stage. The classification stage contains 2 fully connected layers, and softmax output layer. C3D uses convolution filters with size of $k \times k \times d$, where d is the temporal depth and k is the spatial size, with $d = 3$ and $k = 3$. The first pooling layer is defined with size of $2 \times 2 \times 1$ and stride $2 \times 2 \times 1$, whereas the subsequent ones use size of $2 \times 2 \times 2$ with stride $1 \times 1 \times 1$. In the classification stage, the two fully connected layers use 4096 neurons. As already mentioned, we apply transfer learning process to employ a pre-trained C3D on depression estimation. In this process, the fully connected layers are replaced by the ones with 512 neurons, and the softmax layer is replaced by a regression output layer which produces a depression score. This layer maps the features of the last fully connected layer to an output score by:

$$p = \sum_{i=1}^{N} w_i f_i + b, \tag{1}$$

where p is the estimated score, f is the input feature of length N, and w (weight) and b (bias) are the parameters of the regression layer. Fig. 5 illustrates the C3D network used to model spatiotemporal information for depression estimation.

3.2.2 3D global averaging pooling

Given that C3D is a deep model with a high number of trainable parameters, it is important to elaborate schemes to minimize the risk of overfitting. Our proposal is

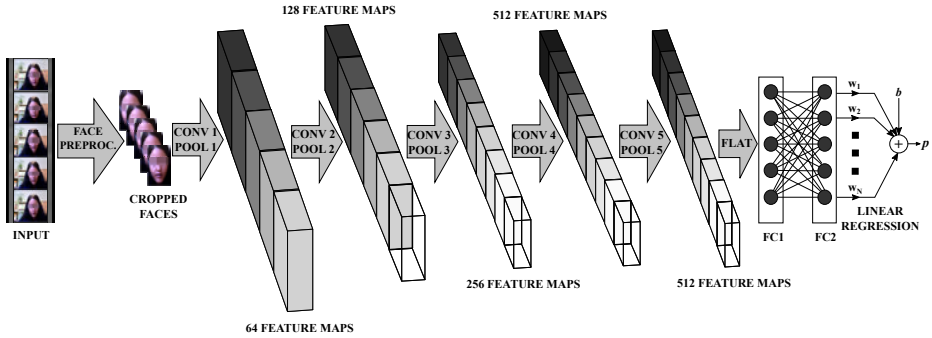


Fig. 5. Illustration of the convolutional 3D network. The preprocessing stage detects and crops faces from video frames. The feature extraction stage employs eight convolutional layers (CONV3, CONV4, and CONV5 have two convolutions) and the classification stage uses two fully connected layers and one regression layer. The output of the network is a depression score. Reprinted, with permission, from Paper I ©2019 IEEE.

to integrate 3D Global Average Pooling (3D GAP) into C3D networks in order to summarize high-level spatiotemporal features from the last convolutional layer. In this process, the last two fully connected layers are replaced by the 3D GAP method which averages features from the last convolutional layer to generate a feature vector that is fed into the regression layer. Specifically, 3D GAP produces a feature vector with 512 values since the last convolutional layer outputs 512 three dimensional features. Such an approach allows a decrease in the number of model parameters leading to reduced time complexity, and better generalization.

The operation performed by 3D GAP may be formally written as:

$$f_n = \frac{1}{Z \times Y \times X} \sum_{z=0}^{Z-1} \sum_{y=0}^{Y-1} \sum_{x=0}^{X-1} H_n(z,y,x), \tag{2}$$

where f_n represents the output feature vector, H_n denotes the three dimensional set of features, $n = 1,2,...,N$, and N is the number of filters in the last convolutional layer. The proposed C3D network with 3D GAP is illustrated in Fig. 6.

3.2.3 Combining global and local C3D networks

The proposed architecture consists of two C3D networks: a local network to explore a facial region, and a global network to explore the full facial region. Considering that the eye region conveys essential information for facial expression recognition [113], we employ the local C3D network to explore a coarse eye region. With that, the architecture pays more attention to the variations of this facial region, which may contribute to the

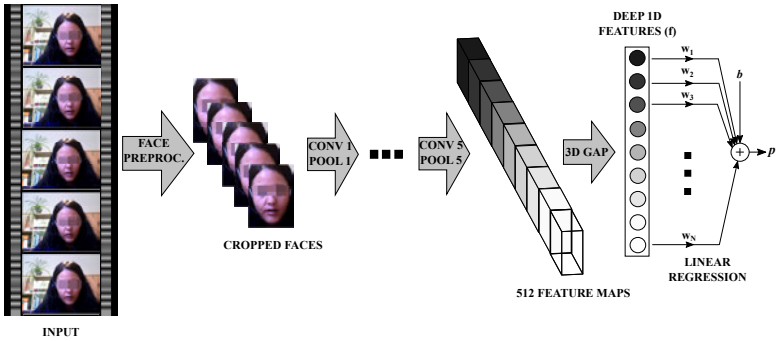


Fig. 6. Illustration of the convolutional 3D network equipped with the 3D GAP method. The spatiotemporal features of the last convolutional layer are summarized by 3D GAP. The resulting features are fed into the regression layer which generates a depression score. Reprinted, with permission, from Paper I ©2019 IEEE.

learning of discriminative features. In addition, the modeling of two facial regions allows the exploration of diverse and complementary features, which may increase the robustness of the learned representations.

In Fig. 7, we illustrate the proposed architecture for the estimation of depression scores. We analyze an input video using an overlapping sliding window (i.e., every video is divided into overlapped clips). The C3D networks explore facial regions that are extracted in the preprocessing stage. A score-level fusion scheme is employed to combine estimations from local and global C3D networks. The scheme uses an operation given by:

$$s_i = p_{li} + |p_{li} + p_{gi}| \times 0.5, \tag{3}$$

where s_i is the depression score for ith clip, p_{li} is the score estimated by the local C3D network, p_{gi} is the score estimated by the global C3D network. To determine the final depression score for an input video, the architecture outputs a depression score for every clip, and computes the average of these values with:

$$S = \frac{1}{I}\sum_{i=1}^{I} s_i = \frac{1}{I}\sum_{i=1}^{I} p_{li} + |p_{li} + p_{gi}| \times 0.5, \tag{4}$$

where I is the total number of clips in the input video.

3.3 Multiscale spatiotemporal network

The previously proposed architecture estimates depression scores using C3D to model facial expression variations. C3D employs filters with size of $3 \times 3 \times 3$ in all convolutional layers to generate representations. Such a homogeneous approach may impair the

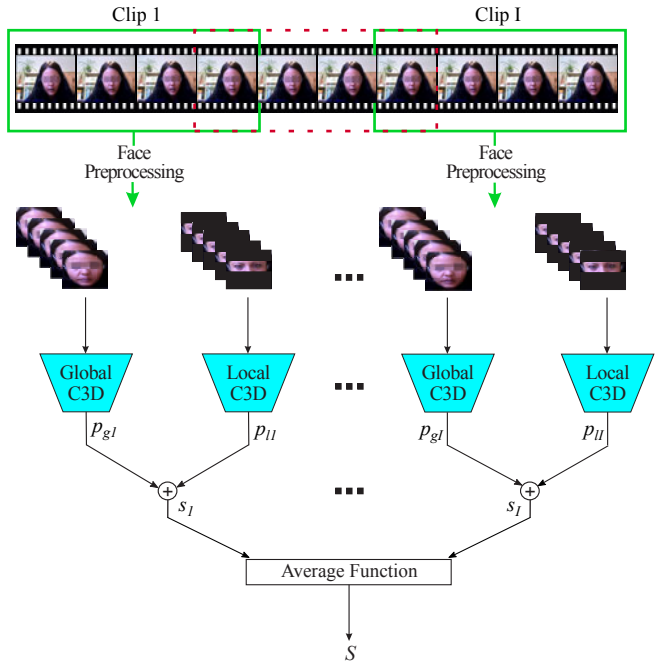


Fig. 7. Illustration of the proposed architecture for depression estimation. Global C3D explores spatiotemporal information from full facial region. Local C3D explores spatiotemporal information from a coarse eye region. A fusion scheme combines the estimations of global C3D (p_g) and local C3D (p_l). The final estimated score (S) is the average of the estimations (output of the fusion scheme) for every clip analyzed from the input video. Reprinted, with permission, from Paper I ©2019 IEEE.

capturing of distinct changes in the facial expressions and the exploration of different facial regions. This limitation affects the extraction of patterns that allow the recognition of different depression levels because the manifestations of depression may involve facial dynamics that consist of a wide range of temporal information and different facial regions convey diverse information about depression levels. Besides the C3D network, other methods have been presented to learn spatiotemporal features. For instance, Carreira *et al.* [99] proposed the Inflated 3D (I3D) network by inflating all the convolutional filters and pooling kernels of 2D Inception model into 3D structures. Hara *et al.* [116] investigated 3D CNNs based on residual connections, called 3D ResNet. Feichtenhofer *et al.* [117] proposed the SlowFast network which is composed of a slow channel to analyze spatial semantics and a fast channel to model temporal correlations at fine resolution. In summary, these deep models basically employ structures that analyze a fixed spatiotemporal range, which leads to the same problem related to the use of C3D

for depression analysis. To tackle this issue, we develop structures with the ability to explore spatiotemporal features in different ranges.

3.3.1 Multiscale structure

Methods based on 3D CNNs employ 3D convolution filters to explore spatiotemporal dependencies in videos. 3D convolutions are defined by the temporal depth and spatial dimensions. The temporal depth determines the range of temporal information that will be analyzed, and the spatial dimensions determine the size of the area of spatial information that will be explored. To effectively generate depression representations, we design our basic building block using multiple 3D convolutions with different temporal depths and spatial sizes.

In Fig. 8, we illustrate our proposed basic building block. The multiscale structure uses identity shortcut connections to link the input x to the output of the last 3D convolutional layer. The residual connection is adopted because it contributes to the training of deeper models by mitigating overfitting problems. We employ three 3D convolutional layers with different spatial sizes and temporal depths. The 3D filters are defined with depth in the range of $d \in \{d_1, d_2, d_3\}$ and spatial size of $h \times w$, where $h \in \{h_1, h_2, h_3\}$ and $w \in \{w_1, w_2, w_3\}$. The output feature maps of the building block is defined as:

$$\bar{y} = \sigma(BN(\mathscr{H}(x, \{H_i\}_{i=1}^{M})) + x), \tag{5}$$

where $\mathscr{H}(.)$ is the function that learns the residual mapping, BN denotes the batch normalization operation, σ is the Rectified Linear Unit (ReLU) activation function,

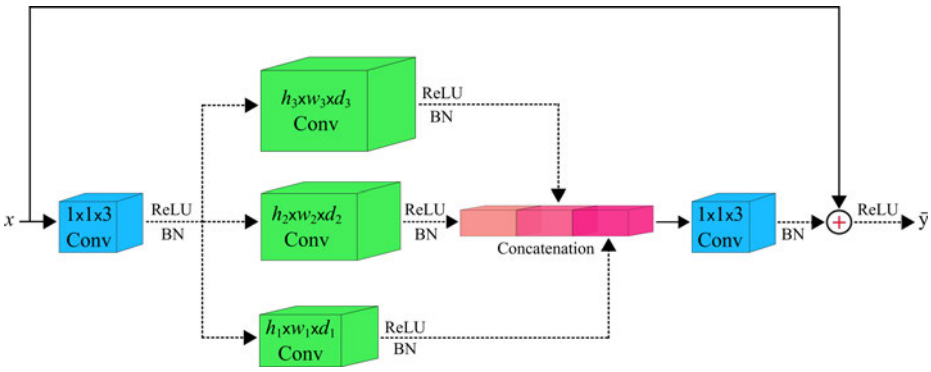


Fig. 8. Illustration of our proposed basic building block. The dashed line indicates an additional operation. BN refers to batch normalization. Reprinted, with permission, from Paper III ©2020 IEEE.

$\{H_i\}_{i=1}^{M}$ denotes the parameters of the convolutional layers, and M is the total number of convolutional layers. The function $\mathscr{H}(.)$ is defined as:

$$\mathscr{H}(x) = H_5\left(\bigcup_{i=2}^{M-1} \sigma(BN(H_i\sigma(BN(H_1x))))\right), \tag{6}$$

where $\bigcup$ denotes the concatenation operation, H_1 represents the parameters of the first convolutional layer, and H_5 denotes the parameters of the last convolutional layer, hence M is equal to 5.

The first convolutional layer (H_1) is defined with convolution kernel of $1 \times 1 \times 3$. This layer receives the feature maps (x) of previous layer and generates output which is fed into three parallel convolutional layers which employ $h_1 \times w_1 \times d_1$, $h_2 \times w_2 \times d_2$, and $h_3 \times w_3 \times d_3$ convolutional filters. We define $h_1 = w_1 = d_1$, $h_2 = w_2 = d_2$, and $h_3 = w_3 = d_3$. To combine the feature maps generated by each parallel convolutional layer, the building block uses a concatenation operation and a $1 \times 1 \times 3$ convolutional layer (H_5). This layer is also used to control the number of channels in the features maps since the concatenation operation increases this number. The first convolutional layer also has this function but it controls the number of channels of x. Hence, H_1 and H_5 layers help to alleviate computational costs. Moreover, ReLU activation function and batch normalization are used after each convolutional layer, with exception of H_5 layer (in this case, only batch normalization is applied).

3.3.2 Multiscale spatiotemporal network

Using the basic building block, we develop a Multiscale Spatiotemporal Network (MSN) to estimate depression scores of individuals in videos. In Fig. 9, we depict an overview of the proposed MSN architecture. The first element of MSN is a $h \times w \times d$ convolutional layer which is defined by $h = w = d = 7$. This layer performs convolutions with spatial stride of 2 and temporal stride of 1, thus spatially downsampling the input of the architecture by a factor of two. In the sequence, a max-pooling layer with kernel size of $3 \times 3 \times 3$ is used to perform a spatiotemporal downsampling. After that, several instances of the basic building block are incorporated in the architecture. Among some of these blocks, max-pooling layers with kernel size of $2 \times 2 \times 2$ and spatiotemporal stride of 2 are applied to reduce the spatiotemporal dimensions of the feature maps by a factor of 2. The spatiotemporal features are reduced in order to simplify the structure of the basic building block in higher layers of the architecture. We use this strategy to control the computational complexity of the architecture. The building block is

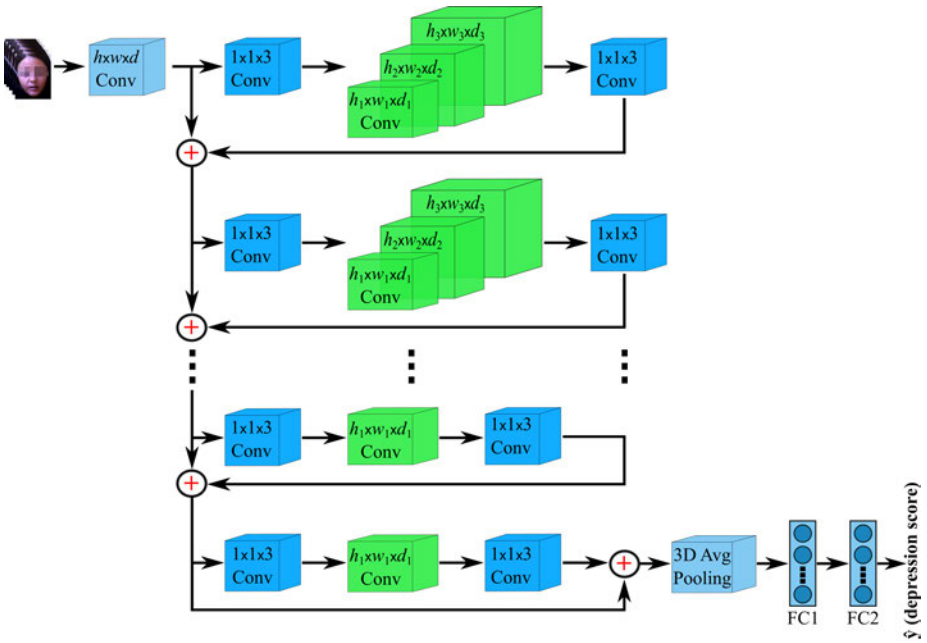


Fig. 9. Illustration of the proposed Multiscale Spatiotemporal Network. The architecture employs 3D parallel convolutions in several layers. In the higher layers, the architecture employs a structure that explores fixed spatiotemporal region. Reprinted, with permission, from Paper III ©2020 IEEE.

simplified by removing the convolutional layers with kernel size of $h_2 \times w_2 \times d_2$ and $h_3 \times w_3 \times d_3$.

MSN analyzes a sequence of facial images to generate a continuous value as output. To optimize the network, we employ Mean Squared Error (MSE) as a regression loss function. Given a training sample n, the MSE is determined by calculating the Euclidean distance between the estimated depression score $\hat{y}_n$ and ground truth value y_n. According to this measure, the loss function of the proposed network is given by:

$$E = \frac{1}{2N} \sum_{n=0}^{N-1} (\hat{y}_n - y_n)^2, \tag{7}$$

where N is the number of samples in the batch (i.e., batch size).

Visualization of activation maps

In order to interpret the capacity of the proposed MSN architecture to explore facial expression variations for depression analysis, we provide Class Activations Maps

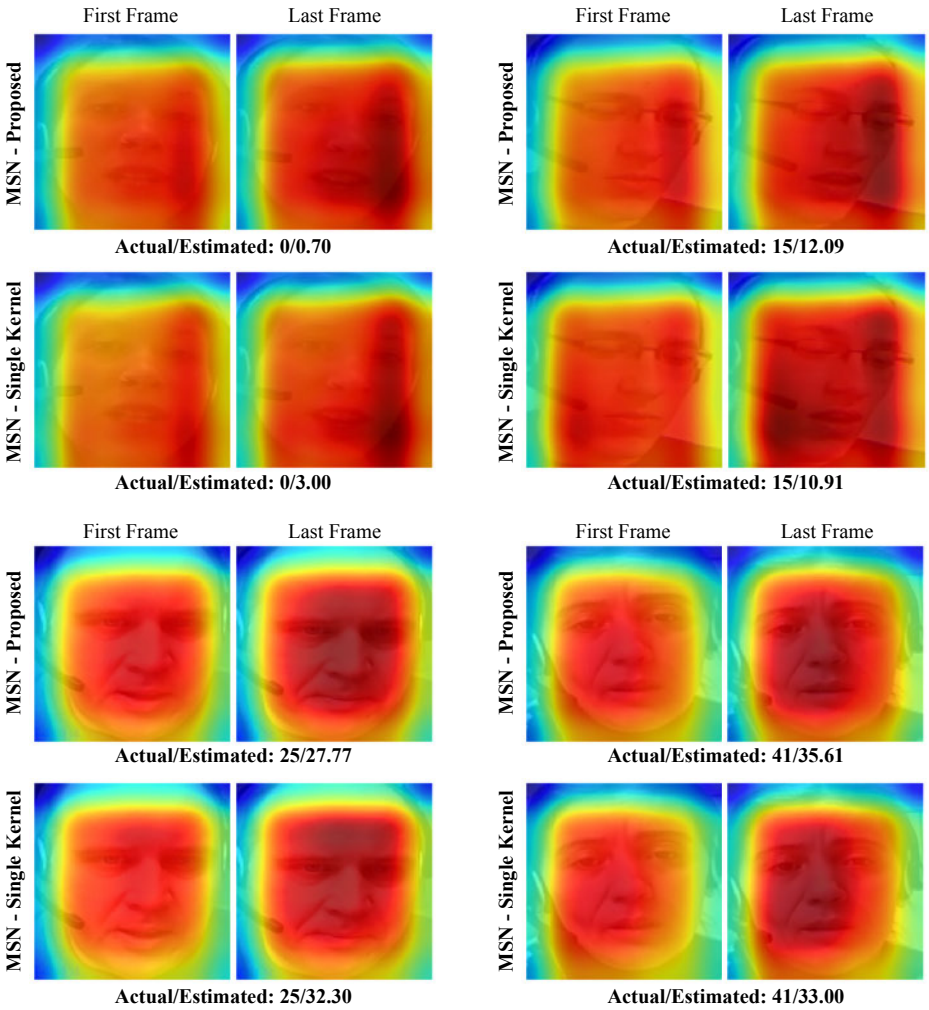


Fig. 10. Visualization of depression attention maps for samples with different depression levels. MSN - Single Kernel refers to the network that uses the basic building block with only one 3D convolutional layer ($h_1 = w_1 = d_1 = 3$). MSN - Proposed refers to the network that uses the basic building block with three parallel 3D convolutional layers. Reprinted, with permission, from Paper III ©2020 IEEE.

(CAMs) using the Grad-CAM algorithm [118]. In the visualizations of Fig. 10, lighter colors describe those regions that are most relevant for the estimation. We show facial activations on two images from an input sample (first and last frames). All depression severity levels are considered in the analysis in order to investigate the spatial and temporal regions that are most informative along the different levels. In general, for all

depression levels, MSN focuses attention on almost the whole facial region, which means that the model captures facial expression variations from diverse facial areas. It is also possible to observe that MSN increases, over time, the region where more attention is paid towards the eye and mouth areas in all depression levels, which also indicates that such areas are important for generating depression representations.

In Fig. 10, we also compare the facial activations produced by MSN with the ones generated by a network with a basic building block using a single kernel (i.e., the $h_2 \times w_2 \times d_2$ and $h_3 \times w_3 \times d_3$ convolutional layers are removed from our proposed block). With that, it is possible to demonstrate the benefits of modeling multiscale spatiotemporal information. Observing facial expression variations in the mouth region, we may see that MSN captures more information from such a region. For instance, examining the facial activations of the individual with minimal level of depression (score equal to 0), it may be noted that MSN intensively explores nearly the entire mouth region, whereas the model with structures using a single kernel has more difficulty to analyze this region. We may also observe a similar result in the individual with a moderate level of depression (score equal to 25). Such results are due to the ability of MSN to investigate a longer range of spatiotemporal variations when compared to the model with structures with a single kernel. Furthermore, MSN seeks to pay attention to the most relevant facial regions. In the example of the individual with mild level of depression (score equal to 15), the model with structures using a single kernel pays high attention to the corner of the face in the first frame, then the exploration of this corner expands, as we can see in the last frame of the sample. Instead of paying high attention to this corner, MSN focuses mainly on facial regions that involve roughly the eyes and mouth regions. For the individual with severe level of depression (score equal to 41), MSN pays high attention to the region involving eyes and mouth which is slightly smaller than the one explored by the model with structures with a single kernel. These two examples indicate that MSN explores more efficiently the spatiotemporal information.

3.4 Depression distribution learning

Deep models extract facial patterns to estimate a depression score for an individual in video. Normally, the estimation of depression scores from facial expressions is considered as a regression problem [98, 102, 103, 104, 105, 106, 110]. It means that only one element of label space is considered to characterize an input. In this scenario, deep models are optimized by using a loss function (e.g., Euclidean loss) to penalize the difference between the estimated score and ground truth. However, these functions are

based on a single depression score (i.e., a real-valued label), so the ordinal relationship between the facial images and depression levels is not explicitly explored. Such an approach has also limited robustness to noisy and uncertain labeling. This limitation may affect the extraction of discriminative features since the depression data is labeled considering the opinion of the patient (e.g., BDI-II questionnaire) or experts (e.g., HAMD-17 instrument), who may have difficulty in evaluating the symptoms. To address these problems, we propose a deep distribution learning method to model the ordinal relationship between the facial images and depression levels. In this way, an instance (e.g., facial image or sequence of images) may be characterized by the intensity (or probability) of each depression score. Our proposed method relies on a new expectation loss to estimate underlying depression distributions. We present an example of distribution estimated for an input image in Fig. 11.

Distribution learning techniques allow the assignment of a label distribution to an object rather than single or multiple labels [119]. A model based on these techniques learns the distribution related to a label space for a sample in order to indicate the level of significance of each label present in this space. During inference, the model may explore this distribution to improve its performance in estimating depression scores. Distribution learning has been successfully used in other image analysis tasks such as age estimation [120] and emotion recognition [121]. For instance, Pan *et al.* [120] proposed to learn distributions for age estimation from faces by using mean-variance loss and softmax loss. However, the methods normally assume that the mean and variance are available during training, which is not the case for depression analysis.

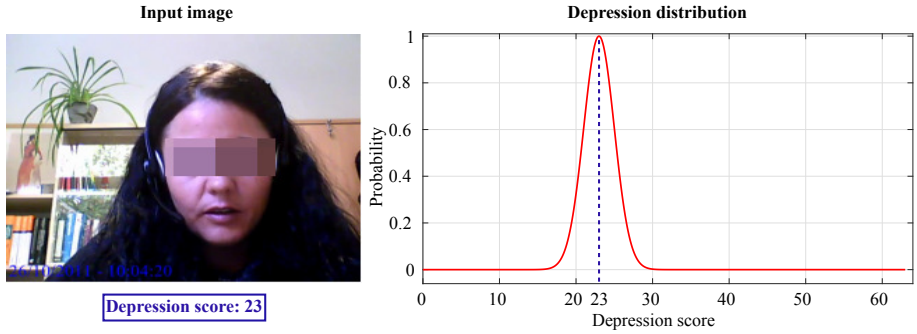


Fig. 11. An example of depression distribution learning. The input is a image with ground truth score equal to 23**. The output is a depression distribution which informs the probability of each depression score for the facial expression in the input image. The distribution is centered at the ground truth of the input (blue line).**

3.4.1 Deep distribution architecture

Our strategy to learn distributions for estimation of depression scores is based on our proposed expectation loss function. The idea is to embed this function into an architecture to learn depression distributions from facial expressions. As shown in Fig. 12, we employ the ResNet-50 model to explore depression distributions for depression analysis. This model uses structures with identity shortcut connections to extract features. Specifically, the basic block is comprised of a stack of convolutional layers with 1×1, 3×3 and 1×1 kernels. The last element of the feature extraction stage is the 2D GAP method that is employed to summarize the features. To generate a depression score, we use a fully connected layer with 512 neurons, softmax layer, and a step to compute the expected values of the depression distribution. Consequently, an output is generated considering the relevance of each depression score in the label space. Since the input of the proposed architecture is a video clip, the final depression score is defined as the average of all estimated values that are computed for each frame. In the following, we describe our expectation loss function.

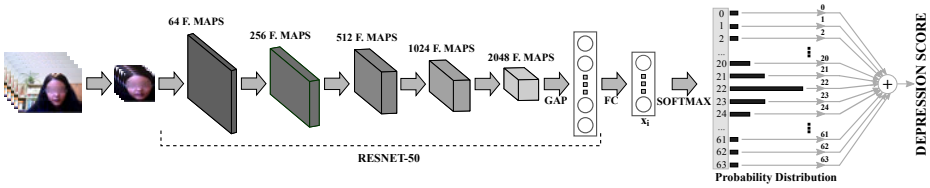


Fig. 12. Illustration of the proposed method for depression distribution learning. ResNet-50 is used as the backbone feature extractor. The softmax arrow represents the softmax layer which is a combination of a fully connected layer with softmax function. Reprinted, with permission, from Paper II ©2020 IEEE.

3.4.2 Expectation loss

To estimate distributions over depression scores, we propose an expectation loss function. This loss technique seeks to penalize the difference between ground truth depression scores and expected values of the depression distributions. Let $y_i \in \{0, 1, .., K-1\}$ represents the depression score label of input i, $\mathbf{x_i}$ denotes the feature representation (see Fig. 12), and $\mathbf{z} = f(\mathbf{x_i}) \in \mathbb{R}^{N \times K}$ refers to the output of the fully connected layer that is employed in the softmax layer. In this case, the softmax probability can be obtained using:

$$p_{i,j} = \frac{e^{z_{i,j}}}{\sum_{m=0}^{K-1} e^{z_{i,m}}}, \tag{8}$$

where $p_{i,j}$ represents the probability that input i belongs to class (depression score label) j, $\mathbf{p_i}$ is the depression distribution estimated for sample i, and $z_{i,j}$ denotes component j of $f(\mathbf{x_i})$. The expected value E_i from the depression distribution $\mathbf{p_i}$ of input i is computed using:

$$E_i = \sum_{j=0}^{K-1} j \cdot p_{i,j}, \tag{9}$$

where j represents depression score labels.

Based on the expected value equation, the expectation loss function of the proposed method for distribution learning is defined as:

$$L = \frac{1}{2M} \sum_{i=0}^{M-1} (E_i - y_i)^2, \tag{10}$$

where M is the batch size. The distance between each expected value and the depression score label is computed using L_2 metric. As a result, we expect that during the training, the model progressively learns a distribution whose expected value approaches the depression score label.

The proposed expectation loss is embedded into the ResNet-50 model for depression estimation in an end-to-end manner (see Fig. 12). Although our architecture is built upon the ResNet-50 backbone, the proposed loss function may be easily incorporated into different deep models, thus helping these models to achieve robust performances.

Visualization of activation maps

In Fig. 13, we present facial attention maps produced by our architecture from samples of each depression severity level. This information allows us to gain insights into the regions that most contribute for depression analysis. The architecture focuses on a region that contains eyes and mouth. We may also observe that the architecture seeks to pay highest attention to a central facial region that includes the mouth area. This result confirms the existent evidences that mouth area conveys important information about depressive states (e.g., less mouth movements [9, 33], shorter smile duration [31, 32], and facial actions [79, 107, 110]: AU 12, AU 14, AU 15, and AU 17). Moreover, partial face occlusion, such as when individuals put their hands on their face, may decrease the performance of the architecture since occlusions hinder the analysis of facial expressions.

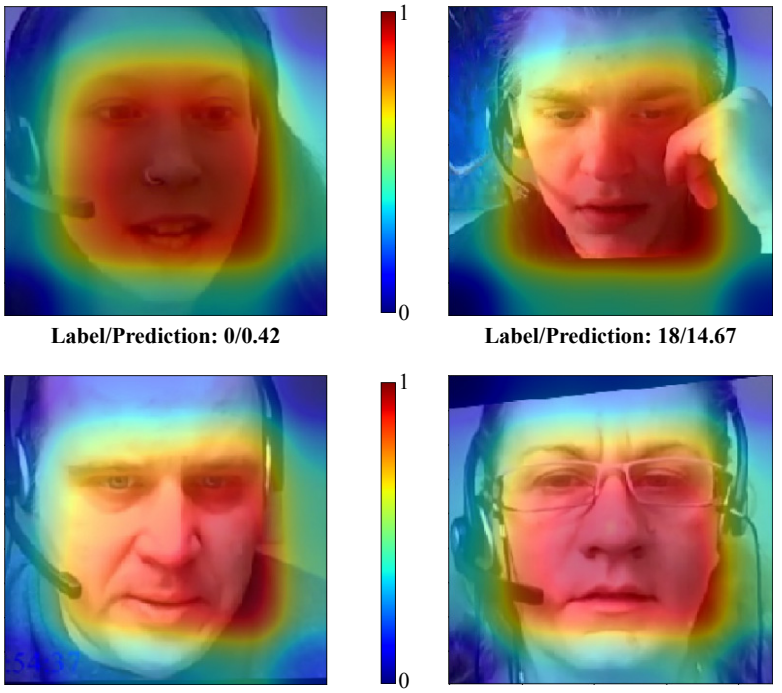


Fig. 13. Facial attention maps for input images with different depression levels. To generate the visualization, we employ the Grad-CAM algorithm [118]. Reprinted, with permission, from Paper II ©2020 IEEE.

3.5 Results and analysis

In order to demonstrate the effectiveness of the proposed methods for depression estimation, we perform extensive experiments on two publicly available depression datasets: AVEC2013 [46] and AVEC2014 [47]. We report the performance of our methods in terms of Mean Absolute Error (MAE) and Root Mean Square Error (RMSE), which are commonly used for depression analysis. The MAE is given by:

$$MAE = \frac{1}{N}\sum_{n=0}^{N-1}|\hat{x}_n - x_n|, \tag{11}$$

where $\hat{x}_n$ denotes the estimated depression score, x_n refers to the ground truth for nth input video, and N is the number of video samples. The RMSE is computed by:

$$RMSE = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1}(\hat{x}_n - x_n)^2}, \tag{12}$$

with identical definitions.

Table 5. Evaluation of components of the global-local network. Local C3D denotes the network that explores local facial region. Global C3D denotes the network that explores full facial region.

Models	AVEC2013		AVEC2014	
	RMSE	MAE	RMSE	MAE
Global C3D	9.30	7.20	8.99	7.23
Local C3D	8.83	6.79	8.87	7.02
Global-local C3D	8.73	6.69	8.76	6.90
Global C3D + 3D GAP	9.24	7.10	8.97	7.09
Local C3D + 3D GAP	8.37	6.51	8.55	6.81
Global-local C3D + 3D GAP	8.26	6.40	8.31	6.59

Global-local network analysis

We begin by investigating the influence of each component of the global-local network on the generation of spatiotemporal representations. This analysis is provided by comparing the performance of the local, global, and global-local C3D networks, with or without the 3D GAP method. In Table 5, we report the results achieved by these networks. As can be seen, local C3D provides the lowest RMSE and MAE values in comparison with global C3D on both depression datasets, with improvement of 5.7% in terms of MAE on AVEC2013 dataset. The use of two C3D networks to explore local and full facial regions (i.e., the global-local C3D network) improves the performance over the C3D networks that explore only one facial region, where the RMSE value is decreased by a margin of 0.23 and 0.11 over global C3D and local C3D on AVEC2014 dataset, respectively. When the 3D GAP method is injected into these networks, we may observe that the networks achieve better results. For example, in terms of RMSE, the performance of the local C3D network is improved by 5.21% on AVEC2013 dataset and 3.61% on AVEC2014 dataset. The boost in performance provided by 3D GAP indicates that the method contributes to reducing the overfitting and improving the generalization capacity of the networks. We may also observe that local C3D demonstrates more potential to explore spatiotemporal variations than global C3D (equipped or not with 3D GAP). It indicates the importance of the spatial correlations and temporal dependencies of a coarse eye region for depression analysis. Moreover, the best performance is achieved by the global-local C3D network equipped with 3D GAP, suggesting that the exploration of

a rich local region combined with a full facial region, that may provide complementary information, favors the modeling of facial expression variations for depression analysis.

Table 6. Evaluation of different configurations for the basic structure of the MSN architecture. Since $h = w = d$, we omit the terms w and d.

3D kernels	AVEC2013		AVEC2014	
	RMSE	MAE	RMSE	MAE
$h_1 = 3$	8.79	6.92	8.36	6.50
$h_1 = 3$ $h_2 = 5$	8.68	6.82	8.14	6.40
$h_1 = 3$ $h_2 = 7$	8.71	6.55	8.45	6.37
$h_1 = 3$ $h_2 = 5$ $h_3 = 7$	7.90	5.98	7.61	5.82
$h_1 = 3$ $h_2 = 5$ $h_3 = 9$	8.18	6.29	7.78	6.04
$h_1 = 3$ $h_2 = 5$ $h_3 = 11$	8.35	6.63	7.74	6.16
$h_1 = 3$ $h_2 = 7$ $h_3 = 9$	8.52	6.96	8.60	6.95
$h_1 = 3$ $h_2 = 7$ $h_3 = 11$	8.82	6.92	8.72	6.76
$h_1 = 3$ $h_2 = 5$ $h_3 = 7$ $h_3 = 9$	8.31	6.51	8.03	6.45

Multiscale ability analysis

We build our MSN architecture using parallel 3D convolutional layers as the basic building block. When we design this structure, three hyperparameters have to be defined: number of parallel 3D convolutions, the temporal depth and spatial size of the filters. These definitions impact the ability to model multiscale spatiotemporal information. To maximize the potential to explore such information, we investigate different configurations for the proposed building block. We analyze structures with up to four parallel 3D convolutions, considering $h_1 = w_1 = d_1$, $h_2 = w_2 = d_2$, $h_3 = w_3 = d_3$, and $h_4 = w_4 = d_4$, i.e., temporal depths and spatial sizes of the 3D convolutional kernels are changed with equal values. In Table 6, we summarize the results of this experiment. It may be observed that the increment of layers for a maximum of three parallel 3D convolutions may contribute to improve the performance of the architecture (four parallel convolutional layers have a high computational cost, which leads to overfitting). However, it is necessary to carefully select the dimensions of the kernels. For example, the model with the basic building block consisting of three parallel layers with $h_1 = 3$, $h_2 = 5$, and $h_3 = 9$ outperforms the one without parallel layers ($h_1 = 3$), whereas the model that uses a structure with $h_1 = 3$, $h_2 = 7$, and $h_3 = 11$ increases the error in relation to the model that employs a structure with $h_1 = 3$. Moreover, considering three

parallel layers, it is possible to observe that the exploration of short, medium and long spatiotemporal ranges favors the performance of the model. An example of this fact may be found in the comparison between the model that uses a structure with $h_1 = 3$, $h_2 = 5$, and $h_3 = 9$ and the one that uses a structure with $h_1 = 3$, $h_2 = 7$ and $h_3 = 11$.

Based on the results from Table 6, we design the basic building block of the MSN architecture using three parallel 3D convolutional layers with $h_1 = 3$, $h_2 = 5$, and $h_3 = 7$ since this structure provides the best performance. Observe that the model built with such a structure outperforms the one that explores a fixed spatiotemporal range ($h_1 = 3$), with improvement of 13.59% in terms of MAE on AVEC2013 dataset and 10.47% in terms of MAE on AVEC2014 dataset. This result confirms our hypothesis that the use of structures with the ability to extract multiscale spatiotemporal features favors the learning of discriminative representations for depression analysis.

Table 7. Evaluation of different loss functions for depression estimation. ResNet-50 is used as backbone of the model.

Method	AVEC2013		AVEC2014	
	RMSE	MAE	RMSE	MAE
Softmax loss	10.22	7.50	10.37	7.89
Euclidean loss	8.48	6.72	8.71	6.37
Expectation loss (ours)	8.25	6.30	8.23	6.15

Expectation loss analysis

Our depression distribution learning approach is based on the proposed expectation loss function. To verify the potential of our loss function, we compare it with two other functions: softmax loss, and Euclidean loss. It is important to note that the softmax loss allows a model to learn a depression distribution, which makes the comparison with this function interesting. The experiments are carried out using the ResNet-50 model to estimate depression scores. We summarize the experimental results in Table 7. As can be seen, the performance of the model significantly improves by using the expectation loss compared to softmax loss. The reason is that the use of only softmax loss during the training phase does not create a sharp probability distribution which impairs the estimation. On the other hand, our expectation loss assists the model, during the training phase, to yield a more concentrated distribution in the direction of the ground truth depression score, which increases the performance of the model. When compared with Euclidean loss, our proposed loss function provides a better performance to the model.

This result indicates that the exploration of the significance of each depression score in the label space may contribute to effectively estimating depression scores.

Table 8. Performance of our proposed methods against other methods on AVEC2013 and AVEC2014 depression datasets. The methods in [46, 47, 88, 90, 92, 94] are based on hand-crafted features. The remaining methods are based on deep features.

Model	AVEC2013		AVEC2014	
	RMSE	MAE	RMSE	MAE
Baseline-AVEC2013 [46]	13.61	10.88	-	-
Baseline-AVEC2014 [47]	-	-	10.86	8.86
MHH+ LBP+ EOH [94]	11.19	9.14	-	-
LGBP-TOP + LPQ [92]	-	-	10.27	8.20
LPQ + SVR [88]	10.82	8.97	-	-
LPQ-TOP [90]	10.27	8.22	-	-
Two-stream network [98]	9.82	7.58	9.55	7.47
DTL [103]	-	-	9.43	7.74
Two 3D CNNs [106]	9.28	7.37	9.20	7.22
ResNet-50 + pooling [104]	-	-	8.43	6.37
Four ResNet-50 [105]	8.28	6.20	8.39	6.21
VGG-16 + FDHH [102]	-	-	8.04	6.68
Global-local C3D (ours)	8.26	6.40	8.31	6.59
ResNet-50 + exp. loss (ours)	8.25	6.30	8.23	6.15
MSN (ours)	7.90	5.98	7.61	5.82

Comparison with state-of-the-art

In Table 8, we show a performance comparison between our proposed methods and the state-of-the-art models for depression estimation on AVEC2013 and AVEC2014 datasets. In general, the models generate depression representations by extracting hand-crafted features or learning deep features. Our proposed methods outperform the models based on hand-crafted features such as LPQ-TOP features [90]. The proposed global-local C3D network achieves comparable results when compared with models based on deep learning techniques. In this regard, the comparison with the model in [106] is interesting because this model uses two C3D to explore full facial area and a region that involves the head. Our global-local C3D outperforms this model by a large margin, which validates our strategy based on the exploration of full facial region and a coarse eye region. A

characteristic of the deep models is that they are based on regression techniques. Our approach based on deep distribution learning, which uses the proposed expectation loss to optimize the ResNet-50 model, achieves better results than almost all deep models. Except for the model in [105] in terms of MAE on AVEC2013 dataset (but this model uses four ResNet-50 to perform estimations) and the model in [102] in terms of RMSE on AVEC2014 dataset. This result further demonstrates the benefit of using our proposed loss function.

From Table 8, it may also be observed that our MSN architecture achieves superior performance over the existing models. It shows that exploring directly the spatiotemporal information across frames (i.e., encoding jointly spatial and temporal information) is a better approach to generate depression representations. However, it is important that the building block of an architecture has the capacity to explore spatiotemporal information at different ranges. Such an ability contributes to the capture of a wide range of facial dynamics and the exploration of different facial regions, allowing a better estimation of depression scores from facial information.

3.6 Summary

The modeling of facial expressions may generate powerful representations for automatic depression analysis. In this process, the analysis of the appearance and dynamic of facial information is essential to learn patterns of depressive behaviors. In other words, spatial and temporal dependencies provide rich information for the extraction of discriminative features. In this sense, we propose two new architectures to effectively explore spatiotemporal information. In the first one, a global and local facial region are explored by two C3D networks. We define a coarse eye region as local and full facial region as global. This approach facilitates the modeling of a highly relevant facial region, i.e., the eye region, and allows the use of the global region to learn complementary features. To improve the ability of generating representations, we integrate the 3D GAP method into C3D. As a result, the architecture yields a better identification of depression scores. In the second one (MSN), a multiscale structure is defined as the basic building block to explore spatiotemporal information at diverse ranges. The basic structure of the architecture is composed of parallel 3D convolutional layers with different temporal depths and spatial sizes. This approach favors the capturing of facial dynamics in different ranges and the exploration of different facial regions. Moreover, we introduce depression distribution learning to analyze depression states. Using this strategy, the model may learn the relevance of each depression score for a sample. To effectively learn depression distributions, we propose the expectation loss function which penalizes

the difference between the expected value of the distribution and ground truth depression score. This approach has the potential to produce robust estimations since it explores the label space to output a depression score.

Our extensive experiments show that the proposed methods have good performance in analyzing facial expressions for depression estimation, indicating that our methods learn effective depression representations. In particular, MSN demonstrates its capacity to generate discriminative representations by outperforming the existing approaches by a large margin. This result indicates that the use of structures with the ability to model multiscale spatiotemporal information is important to overcome the similarity of facial expressions along different depression levels.

Although MSN demonstrates promising results for automatic depression analysis, the key element to achieve such performance is the parallel 3D convolutional layers which have a high computational cost. In the next chapter, we propose two new strategies to reduce computational complexity in the modeling of multiscale spatiotemporal information for depression analysis from facial expressions.

4 Efficient modeling of facial expressions for depression analysis

Automatic depression detection has benefited from deep learning models' ability of extracting depression features from facial expressions. However, the existing depression datasets are limited in size, making it difficult to model facial expressions by a deep approach. Therefore, it is essential to develop deep methods that are able to learn depression patterns from a small amount of training data. In this chapter, we summarize our contributions in Papers IV, V, and VI which present different techniques to efficiently learn depression patterns from facial expressions.

4.1 Introduction

Deep learning techniques have become the solution for many facial analysis applications [38, 39, 40, 41, 122]. These methods automatically analyze facial expressions to learn effective representations. One interesting characteristic of deep methods is that they may be easily adapted to a different application. This aspect is important because usually applications based on facial information (especially healthcare applications) have a limited amount of training data, which hinders the learning process. Indeed, one of the reasons for the success of deep learning is the availability of large amounts of labeled data for applications like image classification [100], so the adaptability of the methods allows the knowledge acquired in one application to be transferred to another application (i.e., transfer learning). This is carried out by initializing the deep models with parameters that were optimized for a task with a large dataset. After the initialization, the deep model may be directly applied in the new task by freezing the parameters in the layers of the feature extraction stage and training a new classification/regression stage. In this case, the model uses features that were learned in the previous task, which may result in an inferior performance. Another option is to fine-tune the model to learn features related to the new task. This may be done by fine-tuning either some or all layers of the feature extraction stage. In both cases, the knowledge of the previous task may contribute to the learning process.

Given the small size of depression datasets, the transfer learning process is commonly applied to improve the extraction of depression patterns. For instance, Al Jazaery *et al.* [106] pre-trained two 3D CNNs on Sports-1M [114] and UCF101 [115] datasets and then fine-tuned all layers of the feature extraction stage on depression datasets.

Since the modeling of facial expressions for depression analysis is a complex process, usually all layers of the feature extraction stage are fine-tuned. This approach may favor the identification of factors of variation in depressive states. Although transfer learning may improve the extraction of depression features from facial information, the use of deep models that have a high computational complexity increases the risk of overfitting. When the model is overfitting, it detects factors that are unimportant to depression analysis (i.e., characteristics of individuals), thus decreasing the ability to extract depression patterns. Moreover, complex deep architectures have a large number of parameters and a high amount of computations, which impact the memory requirements and estimation latency, and make these architectures difficult to deploy on compact platforms with limited resources.

In this chapter, we investigate different deep techniques in order to efficiently learn depression representations. More precisely, we propose a new temporal pooling method and two new architectures that use structures with the ability to explore multiscale spatiotemporal information. The proposed methods are first pre-trained on either VGG Face [123] or VGGFace2 [124] datasets and then fine-tuned on depression datasets. We further employ our two architectures for pain expression recognition.

4.2 Encoding a video segment into 2D representation

The modeling of appearance and dynamic information in video benefits the extraction of depression patterns. To enable CNNs to perform this modeling, the convolution and pooling operations of 2D CNNs are expanded, generating 3D CNNs. Such expansion increases considerably the computational costs, causing an increase in training time and the optimization of a large number of parameters, which makes the deep models prone to overfitting. In order to provide a better alternative to the exploration of facial variations between frames, we propose a temporal pooling method that is employed in a two-stream framework. The proposed approach encodes facial information by capturing and summarizing temporal variations in a sequence of video frames into an image map. The use of this 2D representation allows the exploration of facial dynamics by a 2D CNN, which may be pre-trained on a large dataset and then fine-tuned on depression datasets. This approach favors the learning process since 2D CNNs have lower computational costs compared to 3D CNNs.

The mapping of variations within video into a 2D representation for depression analysis has to be carefully formulated since the changes in facial expressions along the depression levels may be small. A poor representation does not convey discriminative information in its texture, making the exploration of facial dynamics by a 2D CNN

Fig. 14. Example of 2D representations generated by our temporal pooling method. The 2D representations convey dynamic information in their texture, which may be modeled by architectures that explore spatial information.

difficult. Motivated by the success of binary representations for texture analysis [125, 126, 127], we develop our temporal pooling method based on binary code aiming to capture temporal variations, which may help to distinguish depression levels, and facilitate the exploration of movement and velocity in the texture of the representation. Fig. 14 shows examples of representations generated by our proposed method.

4.2.1 The temporal pooling method

Our method analyzes segments of an input video. Suppose $x[m,n,t]$ denotes a segment, where m and n represent the spatial coordinates, and t refers to the temporal coordinate. The first step of the proposed method is to compute the difference around the middle point in the temporal dimension by:

$$z[m,n,t] = \sum_{j=0}^{l-1} x[m,n,j]\delta[m,n,j-t] - x[m,n,(l-1)/2], \tag{13}$$

where $z[m,n,t]$ represents the output signal, $\delta[m,n,t]$ denotes the impulse signal, and l is the amount of frames in the segment. Since the value in the middle point is zero, we simply remove it. To map the direction of the variation, we employ the step activation function defined by:

$$g[m,n,t] = \begin{cases} 0, & \text{if } z[m,n,t] \geq 0 \\ 1, & \text{otherwise.} \end{cases} \tag{14}$$

The resulting signal is encoded by using two operations. In the first, sequences of size α are encoded using XOR operation. In the second, a binary code is employed to

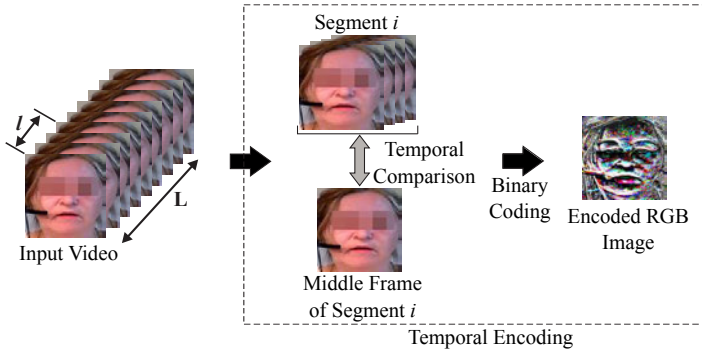


Fig. 15. Overview of the proposed temporal pooling method. An input video is divided into segments. For each segment, a temporal comparison is performed between the middle frame and the other frames of the segment. In the sequence, the method uses a binary code to generate a 2D representation that contains facial dynamic information. Reprinted, with permission, from Paper IV ©2020 IEEE.

generate the final value. The latter operation may be described as follows:

$$y[m,n] = \sum_{j=0}^{h-1} g[m,n,j] \cdot 2^j, \tag{15}$$

where $y[m,n]$ is the 2D representation, which is employed as input of 2D deep models. Specifically, we feed the 2D representations to pre-trained models during the transfer learning process. We define $h = 8$ and the number of frames $l = 2 * \alpha + 1$. In this way, the 2D representation has 8 bits per pixel.

The temporal pooling method is efficient for the learning process in the sense that the method does not employ any trainable parameter. When the method is analyzing an input video, it divides the video into segments, generating several image maps (2D representations) that are used to train a 2D deep model. This is important because a low number of images may lead the deep model to overfitting. In Fig 15, we show an overview of the temporal pooling method.

4.2.2 The two-stream architecture

We insert our temporal pooling method into a two-stream architecture that is composed of an appearance network to explore spatial information and a temporal network to capture dynamic information. The texture of the 2D representations generated by our pooling method is explored by the temporal network. We define the ResNet-50 model as backbone of the temporal network. This model employs identity shortcut connections

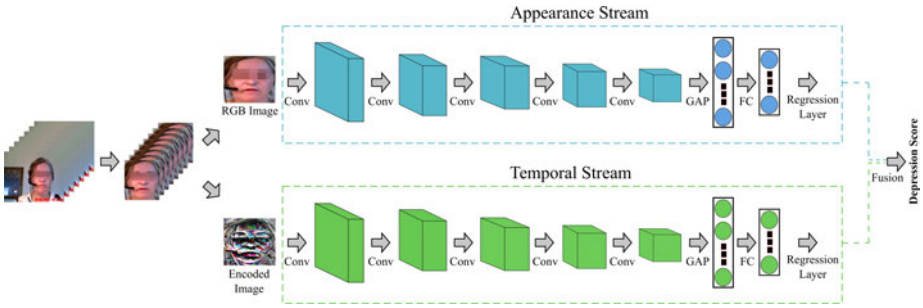


Fig. 16. Illustration of the two-stream architecture. The appearance network explores RGB images to extract features. The temporal network explores the 2D representation generated by our temporal pooling method. The final depression score is produced by the score fusion scheme. Reprinted, with permission, from Paper IV ©2020 IEEE.

and a stack of convolutional layers with kernel sizes of 1×1, 3×3, and 1×1 as basic building block. It also uses the GAP method to summarize the features produced by the last convolutional layer. In the regression stage, we employ a fully connected layer with 512 neurons and a regression layer that computes the estimated depression score.

Given that the facial appearance also conveys valuable static information related to depression, we employ the appearance network to complement the exploration performed by the temporal network. With that, the two-stream architecture captures appearance and dynamic information for depression analysis. Both networks employ ResNet-50 as the backbone, but they are trained separately. We adopt this strategy because the joint training of these models doubles the number of trainable parameters, increasing the overfitting probability. To determine the estimated depression score for an individual, a score fusion scheme averages the scores computed by each network. Fig. 16 illustrates the two-stream architecture equipped with our temporal pooling method.

4.3 The maximization and differentiation network

The capacity of exploring spatiotemporal information is fundamental for the performance of automatic depression detection systems. In Chapter 3, we demonstrated that architectures that employ basic structures with the ability to extract multiscale spatiotemporal features have the potential to generate effective depression representations. In this case, we employ parallel 3D convolutional layers to obtain the capability of multiscale spatiotemporal processing. Such an approach has one disadvantage: it is computationally expensive even in comparison with 3D CNNs.

In order to obtain effective depression representations at reduced computational costs, we propose the Maximization and Differentiation Network (MDN). The basic building block of MDN, namely MDN module, explores spatiotemporal information at different temporal scales by using a maximization block and a difference block. The maximization block is employed to capture smooth spatiotemporal transitions, while the difference block encodes sudden spatiotemporal variations. These blocks do not rely on 3D convolutional layers, which decreases the computing cost, and their learned features are combined into the MDN module in a way that leads to effective representations. The use of this module allows the MDN architecture to encode facial expression variations that are important to depression analysis.

4.3.1 Maximization block

The key idea of the maximization block is to encode global spatial and temporal variations. This may be done by using a function that summarizes such variations in a cascade with 2D convolutional layers that extract relevant spatiotemporal features. Since we design this block to capture smooth spatiotemporal variations of facial expressions, we employ the max function to summarize these variations. The main intuition behind this approach is that the semantic information in an input feature map is redundant along the temporal dimension, which makes it possible to summarize such information without employing trainable parameters. Let $\mathbf{X} \in \mathbb{R}^{N\times T\times H\times W\times C}$ denotes an input feature map, where N,T,H,W and C are the batch size, temporal depth, height, width, and the number of channels, respectively. We formally define the max operation as:

$$\mathbf{V}_{t,h,w} = \max\{\mathbf{X}_{t:t+l,h,w}\}, \tag{16}$$

where $\mathbf{V}$ is the output representation, l is the length of the sliding window employed to perform max pool along the temporal dimension, and t,h,w denote the temporal depth and the spatial dimensions, respectively. Note that the size of the resulting representation is the same as the size of the input feature map.

Instead of exploring spatiotemporal variations with elements that use fixed temporal depths, the maximization block uses a structure that analyzes different ranges of dynamics, contributing to capture multiscale and supplementary information for depression representation. As shown in Fig. 17, the block consists of N branches that operate in distinct ranges, i.e., $l_1, l_2, ..., l_N$. It is worth mentioning that a high number of branches increases the number of parameters, which in turn increases the overfitting probability. On the other hand, a small number of branches may decrease the capabilities of the network. Let $\mathbf{x}^i$ represent the output of branch i, then the block's output may be

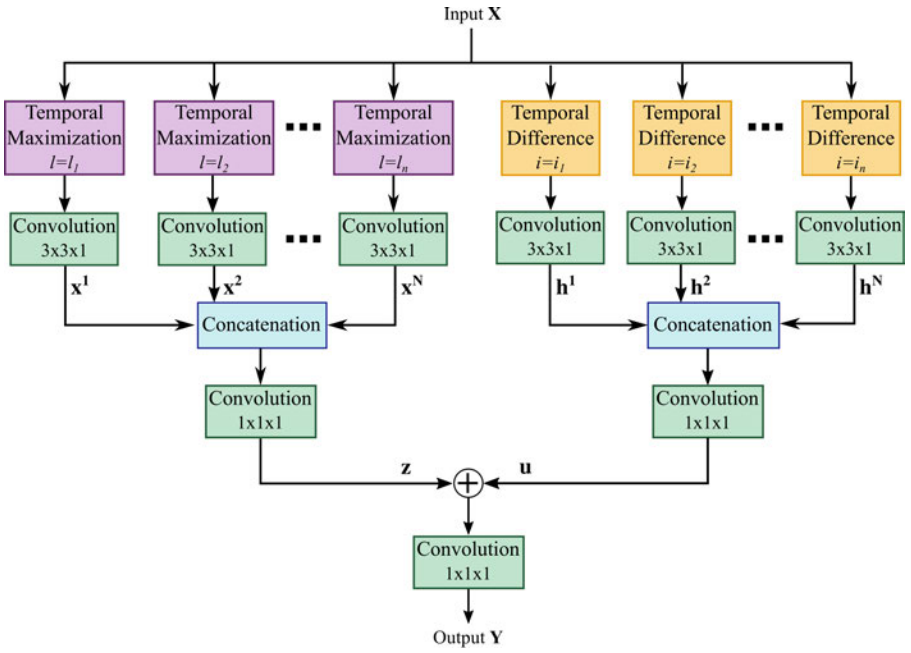


Fig. 17. Illustration of the MDN module. The structure is composed of a maximization block which captures smooth spatiotemporal variations, and a difference block which explores sudden spatiotemporal transitions. A linear combination is carried out at the last stage of the module to generate the output feature map. Reprinted, with permission, from Paper V ©2021 IEEE.

expressed as:

$$\mathbf{z} = \mathscr{H}\left\{\bigcup_{n=1}^{N} \mathbf{x}^n\right\}, \tag{17}$$

where $\mathbf{z}$ denotes the final feature map, $\mathscr{H}\{\}$ is a fusion function performed by a $1 \times 1 \times 1$ convolutional layer, $\bigcup$ is the operation that concatenates the output of each branch, and N denotes the number of branches.

4.3.2 Difference block

The capturing of sudden spatiotemporal transitions of facial expressions may also contribute to generating effective representations. For example, such transitions may assist an architecture to analyze segments of video that have similar facial expressions. From this perspective, we propose a structure called difference block, which explores the velocity of facial expression variations. Considering $\mathbf{X} \in \mathbb{R}^{N \times T \times H \times W \times C}$ as an input

feature map, the first step of the difference block is to compute the absolute value of the difference between the features along the temporal dimension. This operation is defined as:

$$\mathbf{H}_t = |\mathbf{X}_t - \mathbf{X}_{t-i}|, \tag{18}$$

where $\mathbf{H}_t$ is the output of the operation, t is the temporal depth, and i represents ith order difference. In the same way as in the maximization block, the difference block is formed by N branches, which obtain the velocity of the spatiotemporal variations by performing differences of order $i_1, i_2, ..., i_N$ (see Fig. 17). To keep the size of the output representations equal to the size of the input feature map, we add zeros to the input feature map when performing the difference operation.

Since the difference block is designed to explore sudden variations, lower order differences should be employed, such as 1, 2 and 3. The difference block with high order is useful to explore long-term variations. As shown in Fig. 17, a 2D convolutional layer explores the spatial dependencies of the representation generated in the difference operation. We may define the output feature map generated by the difference block as:

$$\mathbf{u} = \mathscr{H}\left\{\bigcup_{n=1}^{N} \mathbf{h}^n\right\}, \tag{19}$$

where $\mathbf{h}^n$ is the output of the convolutional layer of the nth branch, N is the number of branches, $\bigcup$ is the concatenation operator, and $\mathscr{H}$ is the fusion function. Moreover, the operation of concatenation is performed along the channels' dimension in both difference and maximization blocks.

4.3.3 The MDN module

We build our MDN module by combining the maximization block with the difference block. Observe that the maximization block and difference block may operate in different temporal ranges and explore different spatiotemporal information. Consequently, the MDN module has the potential to encode spatial and temporal information from smooth and sudden facial expression variations. Such ability may boost the power to discriminate different depression levels from facial information.

As may be seen from Fig. 17, the output feature maps of the maximization and difference blocks are merged by employing a linear combination. Specifically, the addition operation is used to fuse the features. This operation may only be used if the size of the output feature maps of each block is the same. We use the fusion function $\mathscr{H}$ in the blocks to adjust the size of the feature maps. The advantage of this approach is to

reinforce the complementary behavior of the blocks. The last element of our MDN module is a $1 \times 1 \times 1$ convolutional layer, which is employed to adjust the number of channels of the output to match with the size of the input feature maps ($\mathbf{X}$). With that, the MDN module may be inserted into structures with residual-like connections.

4.3.4 The MDN architecture

The MDN architecture is a convolutional network that generates depression representations by using the MDN module as the basic building block. The architecture consists of four residual layers and one 3D convolutional layer that is used with a different temporal kernel depth to increase the number of temporal ranges explored by the architecture. The residual layers contain a sequence of MDN modules with residual connections. We define the MDN module with a maximization block composed by 3 branches and a difference block formed by 2 branches. All the difference blocks in the architecture perform differences of order 1 and 2 (i.e., $i_1 = 1$, and $i_2 = 2$). Regarding the maximization blocks, we define the operation of the block with temporal depth equal to 2, 3, and 4 (i.e., $l_1 = 2$, $l_2 = 3$, $l_3 = 4$) for the first residual layer. Then, we subtract by one the temporal depth of each branch after every residual layer, where we define the minimum value equal to 1. We adopt this strategy because the spatiotemporal information is downsampled in the second, third, and fourth residual layers. An illustration of the proposed architecture is shown in Fig. 18.

After the residual layers, an average pooling layer with kernel size of $4 \times 4 \times 1$ generates a 256-dimensional feature vector that is fed to the regression stage, which produces a depression score. This stage is composed of a fully connected layer and a linear regression function that we implemented using an additional fully connected layer with one neuron.

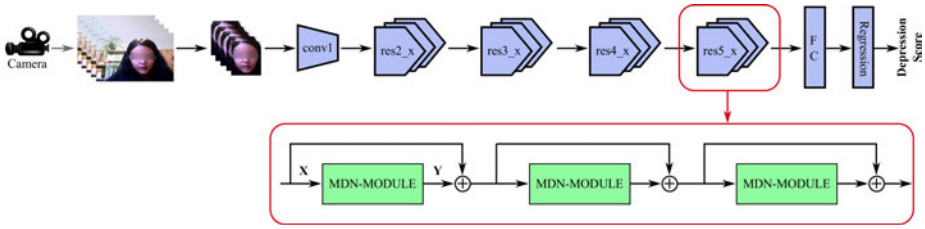
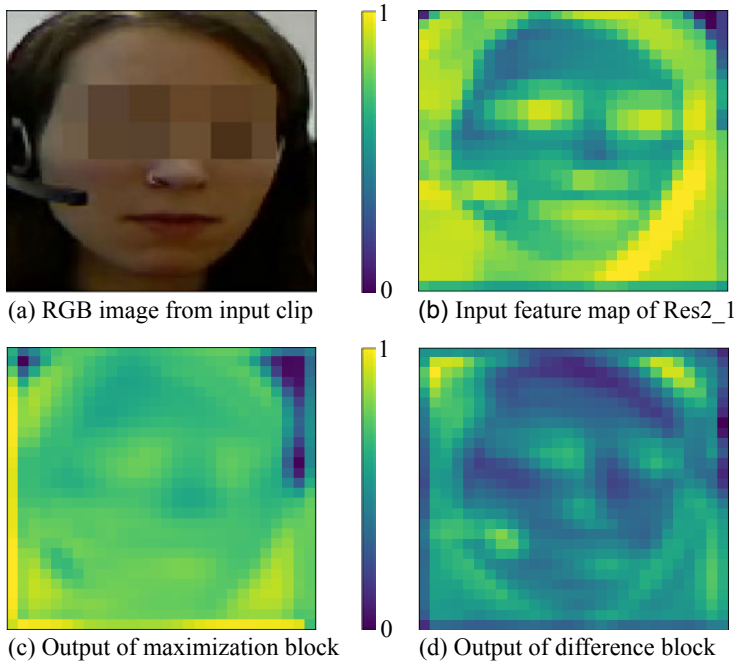


Fig. 18. Illustration of the MDN architecture. Conv1 refers to a 3D convolutional layer. Res denotes the residual layers, which employ a sequence of MDN modules with residual connections. The input of the architecture is a video clip, and the output is a depression score. Reprinted, with permission, from Paper V ©2021 IEEE.

Analysis of feature maps of the MDN module

In order to provide more insight on how the MDN module generates representations, we show in Fig. 19 the depression feature maps generated by the maximization and difference blocks. In our analysis, we consider the first MDN module employed in the first residual layer of the MDN architecture. To facilitate the visualization, we present one frame, Fig. 19(a), from the RGB input clip that is being analyzed by the architecture and the input feature map, Fig. 19(b), of the MDN module. From the output feature map of the maximization block, it may be noticed that this block spreads more energy along the face of the individual when compared to the original input features. It indicates that the block is paying attention to global spatiotemporal information, which increases the potential to explore smooth facial expression variations. From the output feature map of the difference block, it may be seen that the edges and small regions represent the motion captured by this block. Since such a strategy is based on first and second-order differences, the module has the ability to explore sudden spatiotemporal variations. The complementary characteristics of the maximization and difference blocks provides to the MDN module a potential to explore rich spatiotemporal variations.



(a) RGB image from input clip (b) Input feature map of Res2_1

(c) Output of maximization block (d) Output of difference block

Fig. 19. Visualization of feature maps of the MDN module. Reprinted, with permission, from Paper V ©2021 IEEE.

Visualization of activation maps

To visualize the facial regions that are explored by the MDN architecture in order to generate depression scores, we provide the class activation maps using the Grad-CAM method [118]. In the example of Fig. 20, we show facial activation maps generated for four distinct depression levels, which are interpolated and overlaid onto the corresponding facial images. As may be seen, the architecture pays high attention to a region from the eyes to the chin. Interestingly, the most salient area for all depression levels is the region that covers the mouth. We may notice that the different methods proposed in this thesis give high importance to the mouth region, which demonstrates that this region conveys rich information about depression. Furthermore, the visualizations show that our architecture has a different behavior in comparison to the model in [105] which changes the most salient facial region in accordance with the depression level. This fact indicates that the MDN architecture may rely on more optimal facial regions to explore spatiotemporal variations.

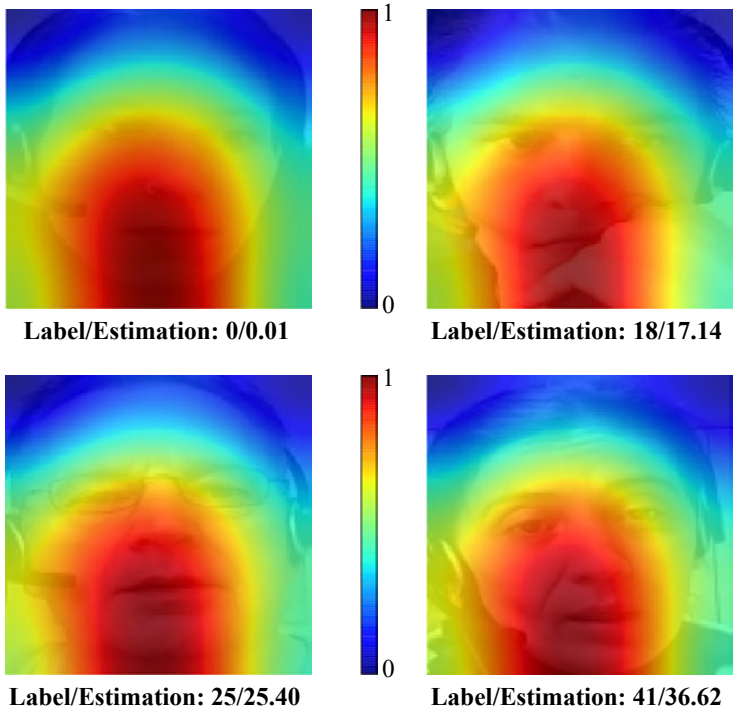


Fig. 20. Facial attention maps for input images with different depression levels. The visualization is generated by employing the Grad-CAM algorithm [118]. Reprinted, with permission, from Paper V ©2021 IEEE.

4.4 The decomposed multiscale spatiotemporal network

To develop an efficient architecture for depression analysis, it is important to consider different aspects of depressive states. One relevant characteristic of depression is its association with pain. Some studies [15, 16] have reported that depressed individuals may suffer from headache, backache, and stomach ache. Therefore, the capability of analyzing facial expressions with characteristics similar to those related to pain may be useful for depression analysis. Studies [51, 128] have associated facial expressions of pain with AU 4 (lowering the brows), AU 6 (cheek raise), AU 7 (lid tightening), AU 9 (nose wrinkling), AU 10 (raising the upper lip), AUs 25-27 (opening the mouth), and AU 43 (eye closure). In summary, a pain event may generate expressions with greater facial variations over time. On the other hand, a depressive state is associated with expressions with fewer variations over time. Fig. 21 illustrates the different facial behaviors of pain and depression. It is also possible to observe that the level of pain may change over time, whereas the depression level lasts for a longer period.

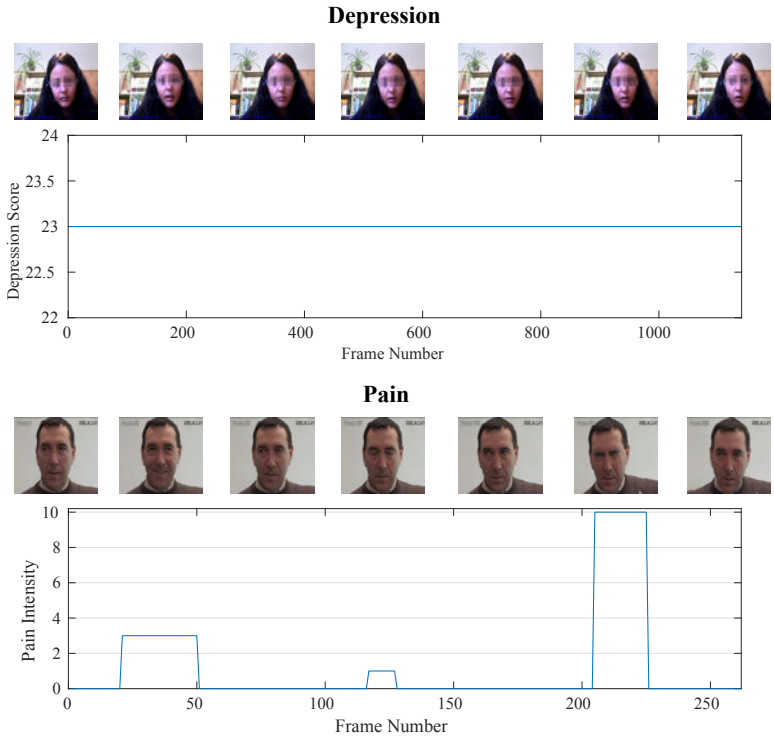


Fig. 21. Examples of depression and pain expressions in sequences of consecutive video frames.

Given that pain and depression have distinct characteristics, an architecture that analyzes both conditions needs to be able to adapt to different facial behaviors. To achieve that, we propose the Decomposed Multiscale Spatiotemporal Network (DMSN). Such an architecture is composed of building blocks that use different strategies to decompose the exploration of multiscale spatiotemporal information in order to reduce computational costs and capture different facial dynamics. The structures are designed employing a sequence of convolutions to increase the range of the facial region in analysis. As shown in Fig. 22, this sequence is called the Main Stage sub-block. The output of each convolution in Main Stage is connected as the input to another convolution that operates in a complementary domain to encode spatiotemporal information. The output feature maps of these branches are at different scales, and a $1 \times 1 \times 1$ convolution combines these features to produce multiscale spatiotemporal representations.

The architectural design of our DMSN building block allows the investigation of different strategies to learn multiscale spatiotemporal features. Since the Main Stage sub-block is responsible for the multiscale ability, it is able to employ convolutions on either the same or different domains, which may be beneficial for the capturing of different facial expression variations. In this context, we develop three variants of our proposed DMSN block (see Fig. 22). In the sequence, we present a detailed description of these variants.

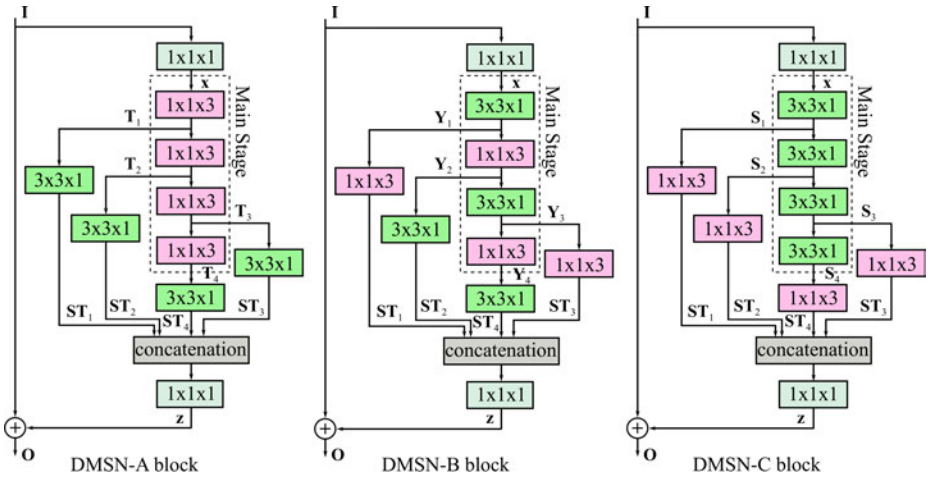


Fig. 22. Illustration of the three proposed DMSN building blocks. Pink blocks denote temporal convolutions, and green blocks denote spatial convolutions. Each convolution in the Main Stage block extends the range of features, and its branches use a complementary convolution to explore spatiotemporal features. These branches generate features that are at different scales, and they are combined using a $1 \times 1 \times 1$ convolution.

4.4.1 DMSN-A block

To develop the DMSN-A block, we consider that 1) the pain intensity level may change more rapidly over time and 2) this level can last for different periods, and 3) pain may produce sudden facial expression variations. For these reasons, we define the Main Stage sub-block of the DMSN-A block as a sequence of $1 \times 1 \times 3$ temporal convolutions. To explore short, medium, and long temporal ranges, this sub-block employs four 1D temporal convolutions. The output $\mathbf{T}_i$ of each 1D temporal convolution ($\mathbf{M}_i^t$) may be computed by:

$$\mathbf{T}_i = \begin{cases} \mathbf{M}_i^t(\mathbf{x}), & i = 1 \\ \mathbf{M}_i^t(\mathbf{T}_{i-1}), & 2 \leq i \leq 4. \end{cases} \tag{20}$$

Each 1D convolution increases the temporal range explored by this sub-block. Branches of the Main Stage sub-block employ $3 \times 3 \times 1$ spatial convolutions which generate spatiotemporal features with fixed spatial size at multiple temporal ranges. The output $\mathbf{ST}_j$ of each 2D spatial convolution ($\mathbf{M}_j^s$) is defined by:

$$\mathbf{ST}_j = \mathbf{M}_j^s(\mathbf{T}_j), \quad 1 \leq j \leq 4. \tag{21}$$

4.4.2 DMSN-B block

The DMSN-B block is developed to learn spatiotemporal features with multiple spatial sizes at multiple temporal ranges. To this end, the Main Stage sub-block is employed to increase the explored regions in both domains by using $3 \times 3 \times 1$ spatial convolutions and $1 \times 1 \times 3$ temporal convolutions. To maintain a similar computational complexity in comparison with the DMSN-A block, we build the DMSN-B block with four convolutions in the Main Stage sub-block. The output $\mathbf{Y}_i$ of each element in this sub-block is calculated by:

$$\mathbf{Y}_i = \begin{cases} \mathbf{M}_i^s(\mathbf{x}), & i = 1 \\ \mathbf{M}_{i-1}^s(\mathbf{Y}_{i-1}), & i = 3 \\ \mathbf{M}_{i/2}^t(\mathbf{Y}_{i-1}), & i = 2,4. \end{cases} \tag{22}$$

Each element of this sub-block increases the spatiotemporal receptive field size in analysis. Branches of the Main Stage sub-block employ complementary convolution (in relation to domain) to generate spatiotemporal features at multiple ranges. Specifically, the output $\mathbf{ST}_j$ of each branch is given by:

$$\mathbf{ST}_j = \begin{cases} \mathbf{M}^t_{j+3-(j+1)/2}(\mathbf{Y}_j), & j = 1,3 \\ \mathbf{M}^s_{j+2-j/2}(\mathbf{Y}_j), & j = 2,4. \end{cases} \tag{23}$$

4.4.3 DMSN-C block

To develop the DMSN-C block, we consider that depressive states may present less facial expression variations over time, and the depression level of an individual in a video tends to be constant. With this in mind, we employ the Main Stage sub-block of the DMSN-C block to generate multiscale spatial features. The sub-block is composed of a sequence of $3 \times 3 \times 1$ spatial convolutions, where each element increases the spatial receptive field size. The output feature map $\mathbf{S}_i$ of each 2D spatial convolution ($\mathbf{M}^s_i$) is computed by:

$$\mathbf{S}_i = \begin{cases} \mathbf{M}^s_i(\mathbf{x}), & i = 1 \\ \mathbf{M}^s_i(\mathbf{S}_{i-1}), & 2 \leq i \leq 4. \end{cases} \tag{24}$$

Branches of the Main Stage sub-block employ $1 \times 1 \times 3$ temporal convolution to generate spatiotemporal features with multiple spatial sizes at a fixed temporal range. The output $\mathbf{ST}_j$ of each 1D temporal convolution ($\mathbf{M}^t_j$) may be given by:

$$\mathbf{ST}_j = \mathbf{M}^t_j(\mathbf{S}_j), \quad 1 \leq j \leq 4. \tag{25}$$

Moreover, the first element of the Main Stage sub-block of the DMSN-A, DMSN-B, and DMSN-C blocks reduces the number of channels by half in comparison with the number of output channels of the first $1 \times 1 \times 1$ convolution, whereas the convolutions in the branches reduces this number by one quarter (i.e., 1 divided by the number of branches).

4.4.4 The DMSN architecture

We build the DMSN architecture by employing our three proposed building blocks (i.e., DMSN-A, DMSN-B, and DMSN-C blocks). The use of our structures allows the exploration of a diversity of multiscale spatiotemporal features, which favors the adaptation to applications with distinct facial behaviors (e.g., pain intensity and depression estimation) and the generation of discriminative representations. The architecture is defined with one convolutional layer (conv1) and four residual layers (res). The residual layers employ a sequence of DMSN blocks. The number of blocks in

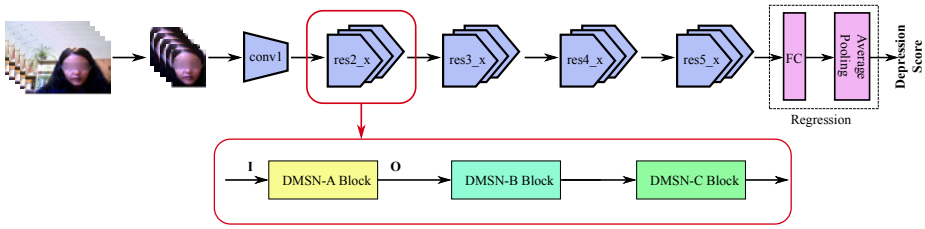


Fig. 23. Illustration of the DMSN architecture. Conv1 represents a 3D convolutional layer. Res represents the layers that employ a sequence of our DMSN blocks. Conv1 and res layers constitute the feature extraction stage. The regression stage is comprised of a fully connected layer and an average pooling operation.

each layer and the model size are defined similarly to the ResNet-50 model. We employ a $3 \times 3 \times 3$ max pooling layer between conv1 and res2 to perform a spatiotemporal downsampling. Conv1 and the first element of each residual layer perform a spatial downsampling. The regression stage is formed by using a fully connected layer and an average pooling operation. In Fig. 23, we show an illustration of the DMSN architecture. Moreover, we develop three models which are named according to the DMSN block they employ. For example, the DMSN-A model employs only the DMSN-A block which explores spatiotemporal features at multiple temporal ranges. With that, we may understand the contributions of each DMSN block for depression analysis.

Visualization of activation maps

To demonstrate the effect of using our three proposed building blocks, we present the class activation maps using the Grad-CAM method [118] for the DMSN architecture, and DMSN-A, DMSN-B, and DMSN-C models. In the visualizations of Fig. 24, lighter colors indicate those facial regions that are most relevant for the estimations of the methods. Analyzing the most activated facial regions, it is possible to observe that the methods appear to explore the eyes and mouth regions. In fact, the eye regions also convey valuable information about depression (e.g., reduced eye contact [36, 37], frowning [34, 35], and facial AU 4 [110]). We may notice that the DMSN architecture and DMSN-C model are more effective to explore these facial regions. Both methods employ the DMSN-C block, which indicates that the strategy of exploring spatiotemporal information at multiple spatial sizes favors the capturing of facial expression variations for depression analysis. In comparison with the DMSN-C model, the DMSN architecture seems to explore these facial regions more intensively. We understand that the diversity of multiscale spatiotemporal features explored by our architecture contributes to this behavior.

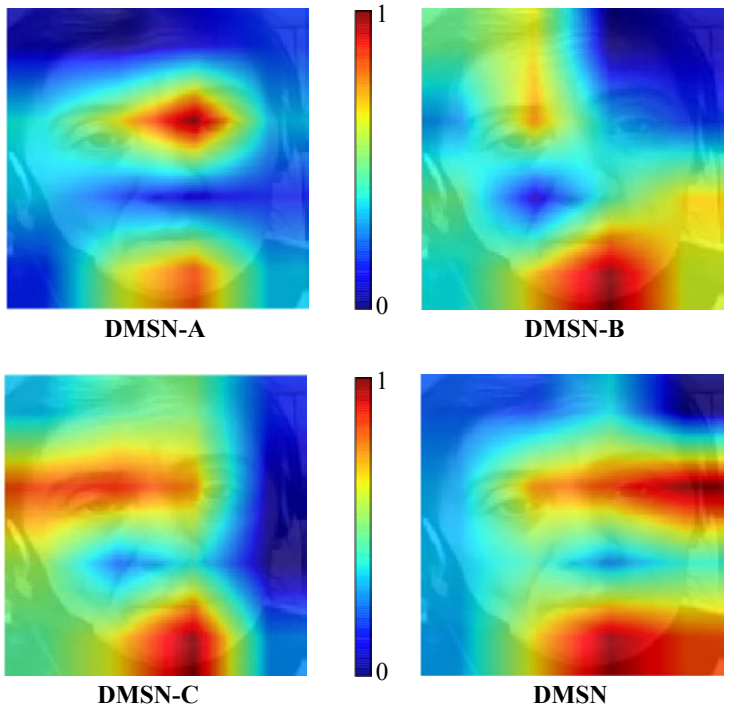


Fig. 24. Facial attention maps of the DMSN architecture, as well as the DMSN-A, DMSN-B, and DMSN-C models.

4.5 Results and analysis

In order to validate our proposed methods for efficient depression representation learning, we carry out experiments on AVEC2013 and AVEC2014 depression datasets and compare the performance of our approaches with the state-of-the-art methods. Moreover, we conduct experiments to evaluate the performance of the MDN and DMSN architectures on the pain intensity estimation task. We decide to analyze MDN for pain estimation because this architecture contains structures to capture sudden facial expression variations.

Two-stream network analysis

In our initial experiments, we analyze the components of the two-stream network (i.e., appearance and temporal networks). We are interested in evaluating the performance of the temporal network that explores the 2D representation (image map) generated by our temporal pooling method. In Table 9, we report the results of both networks separately

Table 9. Evaluation of the different streams in terms of RMSE and MAE for depression estimation.

Method	AVEC2013		AVEC2014	
	RMSE	MAE	RMSE	MAE
Appearance network	8.18	6.15	8.08	6.30
Temporal network	8.43	6.51	8.24	6.29
Two-stream network	7.97	5.96	7.94	6.20

as well as their combination as a two-stream network. The performance achieved by the temporal network is comparable with the one from the appearance network, where the temporal network obtains better results in terms of MAE on AVEC2014 dataset. These results indicate that our temporal pooling method can capture and summarize the dynamic information of facial expressions in video into a 2D representation in such a way that favors the exploration of dependencies of the image. Moreover, the two-stream network outperforms both individual networks. Although the use of two networks (i.e., two ResNet-50) is not computationally cheap, the transfer learning process is performed by training the networks separately, which decreases the chance of overfitting and contributes to efficiently learning depression patterns.

MDN module analysis

The MDN module, which is the basic building block of the MDN architecture, is composed of three branches to explore smooth spatiotemporal variations and two branches to capture sudden spatiotemporal variations. However, the MDN module may be configured with a different number of branches and temporal ranges in both maximization and difference blocks. To study the effect of different configurations, we build MDN models with fifty layers and perform experiments on AVEC2014 dataset. We define the depth (i.e., the length of the sliding window) of the maximization block in range of $1 \leq l \leq 4$, where the temporal depth of the input feature map is considered to define the values of depth in each layer of the model. Regarding the difference block, we define the order values by $i_n = n$, where n is the nth branch. Table 10 summarizes the results of these models that use different configurations for the MDN module. The first model uses MDN modules without the difference block, whereas the second one employs MDN modules without the maximization block. These models achieve similar performance. The third model employs MDN modules that use the same configurations for maximization blocks as the first model and for difference blocks as the second model.

Table 10. Evaluation of MDN models with different configurations for the MDN module. Dep. is the depth value employed in the maximization block, and Order refers to the order of the difference block. Values of Dep. and Order are detailed for each layer of the MDN models. The number of values in Dep. or Order indicates the number of branches. For example, the Order = 1 2 **means that there are two branches in the difference block.**

Layer								AVEC2014	
res2_x		res3_x		res4_x		res5_x			
Dep.	Order	Dep.	Order	Dep.	Order	Dep.	Order	RMSE	MAE
4	-	3	-	2	-	1	-	9.44	7.85
-	1	-	1	-	1	-	1	9.52	7.96
4	1	3	1	2	1	1	1	8.98	7.10
3	1	2	1	1	1	1	1	9.32	7.48
2	1	1	1	1	1	1	1	9.64	7.10
4	1 2	3	1 2	2	1 2	1	1 2	8.64	6.70
3	1 2	2	1 2	1	1 2	1	1 2	9.36	7.37
2	1 2	1	1 2	1	1 2	1	1 2	9.20	7.08
3 4	1	2 3	1	1 2	1	1 1	1	9.00	6.92
2 3	1	1 2	1	1 1	1	1 1	1	8.75	6.71
2 3	1 2	1 2	1 2	1 1	1 2	1 1	1 2	8.40	6.53
3 4	1 2	2 3	1 2	1 2	1 2	1 1	1 2	8.37	6.58
2 3	1 2 3	1 2	1 2 3	1 1	1 2 3	1 1	1 2 3	8.35	6.41
2 3 4	1 2	1 2 3	1 2	1 1 2	1 2	1 1 1	1 2	8.16	6.45

Such a model outperforms the first and second models, demonstrating the importance of exploring smooth and sudden information.

When we compare the models that employ MDN modules composed of difference blocks that use order equals to 1 and maximization blocks with one branch, the model that uses the sequence of 4 (res2), 3 (res3), 2 (res4), and 1 (res5) as depth values achieves a better performance. Employing this sequence of depth values seems to contribute to improving the performance of the model. Normally, increasing the number of branches in the MDN module improves the results of the models. However, the value of depth and order should be carefully selected. For example, the model using MDN modules with values of depth equal to 1 and 2 in all layers and the sequence of 4 (res2), 3 (res3), 2 (res4) and 1 (res5) as depth values outperforms all models that use MDN modules with just two branches, one for maximization block and other for difference block, but this is not true for all models that use MDN modules with two branches for the difference block and one for the maximization block. Comparing the results achieved by models that uses the MDN module formed by two maximization blocks and one difference

block with the models that employ the MDN module composed by one maximization block and two difference blocks, we may note that these models achieve competitive performance. Similar findings may be observed when we compare the model that uses the MDN module composed of three branches for the maximization block and two for the difference block with the model that employs the MDN module with two branches for the maximization block and three for the difference block.

From Table 10, we may also see that the performance of the models that employ the MDN module with two branches for the maximization/difference block and three branches for the difference/maximization block is similar to the models that uses the MDN module with four branches, two for each block. Furthermore, the model that uses the MDN module with values of order equal to 1 and 2 and a maximization block with three branches achieves the best result in terms of RMSE. Based on these results, we decide to specify our MDN module with two branches in the difference block (Order = $[1\ 2]$) and three branches in the maximization block (see the last element of Table 10).

DMSN blocks analysis

The DMSN architecture is designed using DMSN-A, DMSN-B, and DMSN-C blocks. To investigate the potential of each block, we conduct experiments using three models that are named according to the DMSN block used by the model (e.g., DMSN-A only uses DMSN-A blocks). We also compare these models with the DMSN architecture to elucidate the benefits of using all DMSN blocks. The comparison is performed in terms of performance (MAE and RMSE) and memory complexity (number of parameters). In this way, we may evaluate the efficiency of the models. Table 11 reports the results for the three models (i.e., DMSN-A, DMSN-B, and DMSN-C) on AVEC2013 and AVEC2014 datasets. When compared with DMSN-A, DMSN-B achieves better performance, except for AVEC2013 in terms of MAE. DMSN-C obtains the best performance in comparison with DMSN-A and DMSN-B. In terms of memory complexity, it may be seen that DMSN-A employs fewer parameters, whereas DMSN-C has more parameters in comparison with DMSN-A and DMSN-B. Among the three models, DMSN-C provides the best trade-off between performance and memory complexity since this model improves the results with slightly more resources. Based on these results, it is possible to claim that the DMSN-C block is effective to model facial expressions for depression estimation.

Table 11 also shows the performance of our proposed DMSN architecture in comparison with the DMSN-A, DMSN-B, and DMSN-C models. The employment of our three blocks in our architecture provides an improvement of performance over

Table 11. Comparative analysis of the DMSN architecture against DMSN-A, DMSN-B, and DMSN-C models on AVEC2013 and AVEC2014 depression datasets.

Model	AVEC2013		AVEC2014		Parameters
	RMSE	MAE	RMSE	MAE	
DMSN-A	7.98	6.32	8.13	6.48	19.0M
DMSN-B	7.92	6.59	7.86	6.24	23.6M
DMSN-C	7.77	6.14	7.66	6.10	25.9M
DMSN	7.66	6.14	7.50	5.69	22.1M

DMSN-A, DMSN-B, and DMSN-C models (except for AVEC2013 dataset in terms of MAE where DMSN-C obtains the same result). We may observe that the DMSN architecture has lower memory complexity than the DMSN-C and DMSN-B models. Although our architecture has more parameters than DMSN-A, DMSN significantly improves the performance on depression estimation when compared with this model. These results demonstrate that the diversity of multiscale spatiotemporal features extracted by our DMSN architecture enhances the representation for the recognition of depressive states.

Comparison with state-of-the-art

Table 12 presents a comparison in terms of performance and memory complexity between our three proposed methods and the state-of-the-art models for depression estimation on AVEC2013 and AVEC2014 depression datasets. As may be seen, the proposed methods outperform the models based on hand-crafted features by a large margin. Our proposed methods also achieve better results than the models based on deep learning. For instance, our two-stream network, which explores appearance information and the 2D representation generated by our temporal pooling method, outperforms the model in [102], which uses VGG-16 to explore static information and FDHH to capture temporal information. Regarding the memory complexity, the DMSN architecture employs fewer parameters in comparison with other deep models, whereas our two-stream network and the MDN architecture uses fewer parameters than the models in [102, 105, 106]. As indicated by the results, our proposed methods have good potential to infer depressive states from facial expressions.

We also compare our proposed methods with the MSN architecture, which was introduced in Chapter 3, since it achieves a good level of performance for depression estimation. We may observe that our two-stream network and MDN architecture achieve

Table 12. Comparative analysis of our proposed methods against other methods on AVEC2013 and AVEC2014 depression datasets. The methods in [46, 47, 88, 90, 92, 94] are based on hand-crafted features. The remaining methods are based on deep features.

Model	AVEC2013		AVEC2014		Parameters
	RMSE	MAE	RMSE	MAE	
Baseline-AVEC2013 [46]	13.61	10.88	-	-	-
Baseline-AVEC2014 [47]	-	-	10.86	8.86	-
MHH+ LBP+ EOH [94]	11.19	9.14	-	-	-
LGBP-TOP + LPQ [92]	-	-	10.27	8.20	-
LPQ + SVR [88]	10.82	8.97	-	-	-
LPQ-TOP [90]	10.27	8.22	-	-	-
Two-stream network [98]	9.82	7.58	9.55	7.47	-
DTL [103]	-	-	9.43	7.74	-
Two 3D CNNs [106]	9.28	7.37	9.20	7.22	$\approx$ 64.2M
ResNet-50 + pooling [104]	-	-	8.43	6.37	$\approx$ 23.5M
Four ResNet-50 [105]	8.28	6.20	8.39	6.21	$\approx$ 94.0M
VGG-16 + FDHH [102]	-	-	8.04	6.68	$\approx$ 138.0M
MSN (ours)	7.90	5.98	7.61	5.82	77.7M
Two-stream network (ours)	7.97	5.96	7.94	6.20	47.0M
MDN (ours)	7.55	6.24	7.65	6.06	52.0M
DMSN (ours)	7.66	6.14	7.50	5.69	22.1M

competitive results on AVEC2013, but MSN outperforms both methods on AVEC2014. However, our MDN and two-stream network reduce the number of parameters by more than 20M compared with MSN. The DMSN architecture achieves better results than MSN (except for AVEC2013 in terms of MAE), while requiring significantly fewer parameters. These results validate our strategy to decompose the exploration of multiscale spatiotemporal features and the use of a diversity of such features. In short, the DMSN architecture shows the potential to efficiently generate powerful representations for depression estimation.

Pain intensity estimation

To analyze the capacity of our MDN and DMSN architectures to capture facial expressions of pain, we conduct additional experiments on a publicly available dataset called UNBC-McMaster Shoulder Pain Expression Archive. This pain dataset has

Table 13. Comparative analysis of the MDN and DMSN architectures against state-of-the-art methods on UNBC-McMaster pain dataset.

Model	MSE	MAE	Parameters
DCT + LBP [129]	1.39	-	-
HoT [130]	1.21	-	-
OSVR [131]	-	0.81	-
RCNN [132]	1.54	-	-
VGG-11 + LSTM [133]	1.22	0.58	$\approx$ 133.0M
VGG-16 + LSTM [134]	0.74	0.50	$\approx$ 138.0M
C3D [39]	0.71	-	$\approx$ 32.0M
SCN [39]	0.32	-	$\approx$ 586.8M
MDN (ours)	0.68	0.42	52.0M
DMSN (ours)	0.38	0.35	22.1M

been largely used for pain estimation from facial videos. It comprises 200 videos of 25 subjects with a total number of $48,398$ frames. Every video is labeled employing Prkachin and Solomon Pain Intensity (PSPI) scores in a frame-level fashion on a range of 16 discrete levels ranging from 0 (no pain) to 15 (maximum pain). Given that the input of our architectures is a clip, we define as a label the average of the pain intensity of every frame inside the clip. For a fair comparison with other models, we report the performance of MDN and DMSN in terms of Mean Squared Error (MSE) and MAE, where the leave-one-subject-out cross-validation strategy is adopted.

In Table 13, we present the performance of our methods and compare them with the state-of-the-art models. The model in [129] explores Discrete Cosine Transform (DCT) and LBP features. The model in [130] employs Histograms of Topographical (HoT) features, whereas the one in [131] uses Ordinal Support Vector Regression (OSVR) to explore facial landmark points, and LBP and Gabor wavelet features. MDN and DMSN achieve better results than these models. The models in [39, 132, 133, 134] are based on deep learning techniques. In [132], the authors use a Recurrent Convolutional Neural Network (RCNN) as a regression technique to estimate pain levels. Our approaches obtain better performance than this model. The models in [133] and [134] use 2D CNNs to explore spatial information and Long-Short Term Memory (LSTM) to capture temporal information. C3D is employed in [39] to learn spatiotemporal features for pain estimation. We may observe that MDN and DMSN outperform these models, while requiring fewer parameters. The comparison with Spatiotemporal Convolutional Network (SCN) [39] is interesting because the basic building block of this model is

composed of parallel 3D convolutions with diverse temporal depths. MDN and DMSN presents similar performance as this model with significant reduction of parameters, where DMSN has around 26.5 times fewer parameters. These results demonstrate that our architectures are efficient in extracting pain features from facial expressions.

4.6 Summary

Deep learning techniques have the potential to automatically analyze facial expressions to generate representations. One well-known characteristic of these techniques is the high dependence on a large amount of training data. On the other hand, applications based on facial expression analysis normally have a limited amount of data. To overcome this limitation, it is possible to use the knowledge acquired from a large dataset to train a deep model on a small dataset in a process called transfer learning. Automatic depression detection from facial information is an application that has been benefited from deep learning techniques by using transfer learning. However, such an application requires the processing of appearance and dynamics of facial videos, which may lead to the development of complex deep models. This is problematic because these models are prone to overfitting, even if we use transfer learning. Therefore, it is important to devise efficient deep learning techniques that have the potential to learn discriminative representation from a small amount of data. In this context, we introduce a temporal pooling method and two architectures.

The temporal pooling method captures the temporal information in a segment of video to generate a 2D representation. It allows a 2D CNN that was pre-trained on a large dataset to be fine-tuned in order to explore the dynamics of facial expressions for depression analysis. To analyze appearance and dynamic information, we insert the temporal pooling method into a two-stream network, where the temporal network learns depression patterns from the 2D representation and the appearance network explores RGB images to extract depression features. We train these two networks separately to decrease the risk of overfitting and a fusion scheme combines the networks.

MDN is one of the architectures that we propose to efficiently model facial expressions. The basic building block of MDN consists of a maximization block and a difference block. These blocks have different functions. The maximization block is employed to capture smooth facial expression variations, whereas the difference block is responsible to encode sudden facial variations. Both blocks use a function and 2D convolutional layers, thus avoiding the use of expensive 3D convolutions. The combination of these blocks generates the MDN module. Multiple instances of this module are used to form the MDN architecture.

Our last proposed computational model is the DMSN architecture. We design the basic building blocks of DMSN considering the facial expressions of depression and pain. Such an approach may benefit the analysis of depressive states since pain may be one of the symptoms of depression. In total, three blocks are developed, which decompose in different ways the exploration of multiscale spatiotemporal information, resulting in a reduction of computational costs. The use of these blocks also allows the DMSN architecture to explore a variety of multiscale spatiotemporal features.

Our extensive experiments indicate that the representations learned by the proposed methods obtain superior performances over the ones produced by existing models for depression estimation. MDN and DMSN also demonstrate to be an efficient option for pain intensity estimation. Overall, DMSN has good potential to effectively and efficiently model facial expressions, which makes this architecture a cost-effective solution for applications like depression and pain intensity estimation.

5 Conclusion

Depression has been recognized as a serious and common mental health disorder, with an immense economic burden. An accurate clinical evaluation of depression is essential to provide appropriate treatment and reduce overall economic costs. Traditional depression assessments rely on the clinician's understanding of verbal reports from patients or self-report questionnaires. The subjective nature of these methods has resulted in difficulties to evaluate depression. This problem has been propelling the design of automatic depression detection systems to provide objective and reliable medical tools that may assist health professionals. In recent years, advancements have been made towards this goal, especially for systems that analyze facial information. Indeed, there is strong evidence that depressive states modify facial behavior. Diminished facial expressiveness, decreased positive emotional facial expressions, reduced eye contact, and reduced mouth movements are some examples that reflect depression.

The automatic depression detection methods based on facial information may offer contact-free solutions by exploring correlations between facial expressions captured in videos and depressive conditions. Since the analysis of temporal evolution assists to distinguish depressive and healthy states, it is important to build systems that explore spatial and temporal information within facial videos. The exploration of spatial information allows the modeling of facial appearance and the analysis of temporal evolution favors the modeling of facial dynamics. Different methods have been proposed to generate depression representations from facial videos. Conventional approaches use hand-crafted features to represent an input, whereas deep learning techniques automatically analyze the dependencies of the input to generate representations. In comparison with hand-crafted methods, deep learning techniques have shown a higher level of performance.

5.1 Contributions of the thesis

In this thesis, we focused on the development of computational models based on deep learning techniques to explore facial expressions for the problem of estimating a depression score for an individual in video. The proposed methods learn depression patterns in a scenario of intrinsic challenges related to depression (e.g., small differences in facial expressions along different levels) and automatic facial analysis such as large intra-class variations. Our contributions may be divided into two parts: effective

depression representation learning from facial information, and efficient modeling of facial expressions for depression analysis.

Effective Representation Learning. We began by investigating strategies to generate discriminative representations from the appearance and dynamics of facial expressions. Our first architecture employs two C3D networks to explore spatiotemporal information from local and global regions. We defined a coarse eye region as local and the full facial region as global. The estimation of each network is combined by using a score fusion scheme. Despite the interesting potential of this architecture, the use of a basic structure that explores a fixed spatiotemporal range limits the capacity to extract depression features. In the second architecture, we employed a basic building block composed of parallel 3D convolutional layers to explore spatiotemporal information at diverse ranges, favoring the capturing of different ranges of dynamics and the exploration of facial regions that convey essential information about depression. The use of this structure results in an architecture (MSN) that generates effective representations. Moreover, we proposed to formulate the depression estimation problem as a distribution learning problem. In this way, an architecture learns the importance of each element of the label space for an instance. Our approach is based on a new expectation loss function and it has the potential to improve the robustness of the estimations.

Efficient Representation Learning. The modeling of facial expressions using deep learning techniques has shown promising performance in detecting different health conditions. Such techniques require a large amount of training data, but applications like automatic depression detection have datasets that are small in size. To mitigate this problem, transfer learning techniques have been used to improve the learning process. However, the need of spatiotemporal processing of facial videos may lead to complex deep models, which creates difficulties to learn depression patterns. We addressed this problem by proposing a temporal pooling method and the MDN and DMSN architectures. The temporal pooling method generates 2D representations by capturing temporal information within video frames. The representations convey facial dynamic information in their texture which may be encoded by a less complex deep model developed to explore spatial information. Our temporal pooling method is inserted into a two-stream network, where the networks are trained separately and a fusion score scheme combines the networks. Motivated by the promising results of MSN, we developed new basic building blocks to efficiently perform multiscale spatiotemporal processing. First, we designed the MDN module, which is comprised of maximization and difference blocks, to explore smooth and sudden spatiotemporal variations. We used multiple instances of this module to build an architecture (MDN) that has the potential to explore different facial expression variations. In addition, we

designed three blocks that employ different strategies to decompose the exploration of multiscale spatiotemporal features. We observed that exploring the spatiotemporal information at multiple spatial sizes (DMSN-C block) contributes to learning depression patterns. We built our DMSN architecture by employing these blocks, which explore a diversity of multiscale spatiotemporal features and demonstrate to be a cost-effective solution for depression estimation. Both MDN and DMSN architectures also generate efficient representations for pain estimation.

5.2 Future work

This thesis proposed diverse deep learning techniques to model facial expressions for automatic depression detection. The developed methods elucidate the importance of exploring appearance and dynamics of facial videos to generate representations that allow the recognition of different depression levels. Specifically, the multiscale spatiotemporal ability demonstrates to be essential to learn depression patterns. From the perspective of depression analysis, there are some directions that may be taken to further contribute to the advancement of the area. In the following, we discuss three future research directions:

Multi-modal depression analysis

Although facial expressions convey valuable information about depression, other modalities also contain useful information for depression analysis. For instance, speech is considered an important clinical marker of depression [9, 135]. A person suffering from depression is commonly associated with a reduced speech rate [9], speech that has longer pauses [136], and monotonous pitch [137]. Consequently, speech and facial information may be jointly explored to boost the learning of depression features. Normally, this is performed by using one network to explore facial expressions and one network to encode audio information. These networks may be combined at the feature level [102, 138] or score level [92, 139]. However, this approach is suboptimal because one modality may be more discriminating than the other in some videos. A simple example of this case is when there is the occurrence of facial occlusions or a mixture of speech with background noise. Therefore, an effective audio-visual system pays more attention to the modality that is more informative. Moreover, such systems may include other modalities such as text [140], electrodermal activity [141], electroencephalogram (EEG) [142], and heart rate variability [143].

Exploring correlations between depression and other health conditions

There are correlations between depression and other mental health disorders [144, 145, 146, 147]. Studies [144, 148] report that 58% of individuals suffering from depression experience anxiety disorder. There also exists considerable evidence supporting the association between depression and stress [145, 149]. Consequently, a deep architecture could explore the correlations between depression, anxiety, and stress in order to learn common, yet powerful representations. In this case, one feature extraction stage could explore facial expressions of these health states and three different classification/regression stages could generate outputs related to each condition, which constitutes a multi-task approach. Since there is no publicly available large-scale labeled dataset for these healthcare applications, the main advantage of this approach is the augmentation of the labelled training data for each application. Other mental health conditions may also be considered to increase the amount of data, but it is important to select conditions that have a high correlation with depression.

Self-Supervised learning for depression analysis

Collecting data for automatic depression detection is difficult due to privacy and ethical reasons. Another aspect is that the process of labeling facial videos of depressed individuals is expensive and time-consuming. These facts explain the availability of depression datasets that are small in size. As we mentioned previously, this characteristic of depression datasets is a limitation to train deep learning models. Our approach to deal with this problem is to use transfer learning techniques. Recently, self-supervised learning [150] has emerged as an effective approach to handle the training of deep models on small datasets [151, 152, 153]. In this approach, the deep model is first trained on large-scale unlabeled data to solve a pretext task, and then the learned representations are used in a downstream task, where the model is fine-tuned on a labeled dataset. Usually, the pretext task consists of applying a transformation to the input image, and then to estimate the properties of the transformation. Affine transformations [154] and rotations [155] are some examples of transformations. Moreover, the downstream task is an application that normally employs smaller datasets in comparison with the ones used in the pretext task. An investigation of depression estimation as the downstream task could demonstrate that this approach contributes to learning depression features.

5.3 Concluding remarks

The research on automatic depression detection is relatively new and challenging. This thesis takes one important step towards the goal of building an assistive medical tool for depression analysis by developing deep learning techniques to model facial expressions. In summary, we demonstrated that the exploration of appearance and dynamic information contributes to generating discriminative depression representations, especially when such exploration is performed using structures with multiscale spatiotemporal representation ability. Although the proposed methods show promising performance, there is still a need for further research and development in order to reach the point where these solutions may be deployed in real-world scenarios. We hope that the present work provides the basis for future research.

References

[1] World Health Organization, "Depression and other common mental disorders: Global health estimates," Available at https://apps.who.int/iris/bitstream/handle/10665/254610/WHO-MSD-MER-2017.2-eng.pdf, 2017, accessed: 22-11-2021.

[2] S. M. Monroe and K. L. Harkness, "Recurrence in major depression: A conceptual analysis," *Psychological Review*, vol. 118, no. 4, pp. 655–674, 2011.

[3] W. F. Stewart, J. A. Ricci, E. Chee, S. R. Hahn, and D. Morganstein, "Cost of lost productive work time among US workers with depression," *JAMA*, vol. 289, no. 23, pp. 3135–3144, 2003.

[4] A. C. Hummel, E. J. Kiel, and S. Zvirblyte, "Bidirectional effects of positive affect, warmth, and interactions between mothers with and without symptoms of depression and their toddlers," *Journal of Child and Family Studies*, vol. 25, no. 3, pp. 781–789, 2016.

[5] U. S. Rehman, J. Gollan, and A. R. Mortimer, "The marital context of depression: Research, limitations, and new directions," *Clinical Psychology Review*, vol. 28, no. 2, pp. 179–198, 2008.

[6] J. Arias-de la Torre, G. Vilagut, A. Ronaldson, A. Serrano-Blanco, V. Martín, M. Peters, J. M. Valderas, A. Dregan, and J. Alonso, "Prevalence and variability of current depressive disorder in 27 european countries: A population-based study," *The Lancet Public Health*, vol. 6, no. 10, pp. e729–e738, 2021.

[7] D. McDaid, "Making the long-term economic case for investing in mental health to contribute to sustainability," *European Union*, 2011.

[8] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*. American Psychiatric Publishing, 2013.

[9] J. F. Cohn, N. Cummins, J. Epps, R. Goecke, J. Joshi, and S. Scherer, *Multimodal assessment of depression from behavioral signals*. Association for Computing Machinery and Morgan & Claypool, 2018, p. 375–417.

[10] A. Pampouchidou, P. G. Simos, K. Marias, F. Meriaudeau, F. Yang, M. Pediaditis, and M. Tsiknakis, "Automatic assessment of depression based on visual cues: A systematic review," *IEEE Transactions on Affective Computing*, vol. 10, no. 4, pp. 445–470, 2019.

[11] K. Demyttenaere, J. De Fruyt, and S. M. Stahl, "The many faces of fatigue in major depressive disorder," *International Journal of Neuropsychopharmacology*, vol. 8, no. 1, pp. 93–105, 2005.

[12] M. Murphy and M. J. Peterson, "Sleep disturbances in depression," *Sleep Medicine Clinics*, vol. 10, no. 1, pp. 17–23, 2015.

[13] J. S. Buyukdura, S. M. McClintock, and P. E. Croarkin, "Psychomotor retardation in depression: Biological underpinnings, measurement, and treatment," *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, vol. 35, no. 2, pp. 395–409, 2011.

[14] C. Martin-Soelch, "Is depression associated with dysfunction of the central reward system?" *Biochemical Society Transactions*, vol. 37, pp. 313–317, 2009.

[15] S. Borgman, I. Ericsson, E. K. Clausson, and P. Garmy, "The relationship between reported pain and depressive symptoms among adolescents," *The Journal of School Nursing*, vol. 36, no. 2, pp. 87–93, 2020.

[16] S. M. Stahl, "Does depression hurt?" *The Journal of Clinical Psychiatry*, vol. 63, pp. 273–274, 2002.

[17] J. M. Hettema, "What is the genetic relationship between anxiety and depression?" *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, vol. 148C, no. 2, pp. 140–146, 2008.

[18] J. C. Levy and E. Y. Deykin, "Suicidality, depression, and substance abuse in adolescence," *American Journal of Psychiatry*, vol. 146, no. 11, pp. 1462–1467, 1989.

[19] K. Hawton, C. C. i Comabella, C. Haw, and K. Saunders, "Risk factors for suicide in individuals with depression: A systematic review," *Journal of Affective Disorders*, vol. 147, no. 1, pp. 17–28, 2013.

[20] A. McGirr, J. Renaud, M. Seguin, M. Alda, C. Benkelfat, A. Lesage, and G. Turecki, "An examination of dsm-iv depressive symptoms and risk for suicide completion in major depressive disorder: A psychological autopsy study," *Journal of Affective Disorders*, vol. 97, no. 1, pp. 203–209, 2007.

[21] J. L. Sotelo and C. B. Nemeroff, "Depression as a systemic disease," *Personalized Medicine in Psychiatry*, vol. 1-2, pp. 11–25, 2017.

[22] A. J. Mitchell, A. Vaze, and S. Rao, "Clinical diagnosis of depression in primary care: A meta-analysis," *The Lancet*, vol. 374, pp. 609–619, 2009.

[23] M. J. Bostwick, "Recognizing mimics of depression: The '8 ds'," *Current Psychiatry*, vol. 11, no. 6, pp. 31–36, 2012.

[24] K. Kroenke and R. L. Spitzer, "The phq-9: A new depression diagnostic and severity measure," *Psychiatric Annals*, vol. 32, no. 9, pp. 509–515, 2002.

[25] M. Hamilton, "A rating scale for depression," *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 23, pp. 56–62, 1960.

[26] G. Purebl, I. Petrea, L. Shields, M. D. Tóth, A. Székely, T. Kurimay, D. McDaid, E. Arensman, I. Granic, and K. M. Abello, "Depression, suicide prevention and e-health: Situation analysis and recommendations for action," *The Joint Action on Mental Health and Well-Being*, 2015.

[27] B. Renneberg, K. Heyn, R. Gebhard, and S. Bachmann, "Facial expression of emotions in borderline personality disorder and depression," *Journal of Behavior Therapy and Experimental Psychiatry*, vol. 36, no. 3, pp. 183–196, 2005.

[28] D. M. Sloan, M. E. Strauss, and K. L. Wisner, "Diminished response to pleasant stimuli by depressed women," *Journal of Abnormal Psychology*, vol. 110, no. 3, p. 488–493, 2001.

[29] W. Gaebel and W. Wölwer, "Facial expressivity in the course of schizophrenia and depression," *European Archives of Psychiatry and Clinical Neurosciences*, vol. 254, no. 5, p. 335–342, 2004.

[30] H. Berenbaum and T. F. Oltmanns, "Emotional experience and expression in schizophrenia and depression," *Journal of Abnormal Psychology*, vol. 101, no. 1, p. 37–44, 1992.

[31] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, and L.-P. Morency, "Automatic behavior descriptors for psychological disorder analysis," in *Proc. IEEE International Conference on Automatic Face & Gesture Recognition*, 2013, pp. 1–8.

[32] S. Scherer, G. Stratou, G. Lucas, M. Mahmoud, J. Boberg, J. Gratch, A. S. Rizzo, and L.-P. Morency, "Automatic audiovisual behavior descriptors for psychological disorder analysis," *Image and Vision Computing*, vol. 32, no. 10, pp. 648–658, 2014.

[33] J. T. M. Schelde, "Major depression: Behavioral markers of depression and recovery," *The Journal of Nervous & Mental Disease*, vol. 186, no. 3, p. 133–140, 1998.

[34] A. E. Kazdin, R. B. Sherick, K. Esveldt-Dawson, and M. D. Rancurello, "Nonverbal behavior and childhood depression," *Journal of the American Academy of Child Psychiatry*, vol. 24, no. 3, pp. 303–309, 1985.

[35] G. Stratou, S. Scherer, J. Gratch, and L.-P. Morency, "Automatic nonverbal behavior indicators of depression and ptsd: Exploring gender differences," in *Proc. Humaine*

Association Conference on Affective Computing and Intelligent Interaction, 2013, pp. 147–152.

[36] G. M. Lucas, J. Gratch, S. Scherer, J. Boberg, and G. Stratou, "Towards an affective interface for assessment of psychological distress," in *Proc. International Conference on Affective Computing and Intelligent Interaction*, 2015, pp. 539–545.

[37] L. A. Fairbanks, M. T. McGuire, and C. J. Harris, "Nonverbal interaction of patients and therapists during psychiatric interviews," *Journal of Abnormal Psychology*, vol. 91, no. 2, p. 109–119, 1982.

[38] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, pp. 1–20, 2020.

[39] M. Tavakolian and A. Hadid, "A spatiotemporal convolutional neural network for automatic pain intensity estimation from facial dynamics," *International Journal of Computer Vision*, vol. 127, p. 1413–1425, 2019.

[40] M. B. Lopez, C. R. del Blanco, and N. Garcia, "Detecting exercise-induced fatigue using thermal imaging and deep learning," in *Proc. International Conference on Image Processing Theory, Tools and Applications*, 2017, pp. 1–6.

[41] S. Jaiswal, M. F. Valstar, A. Gillott, and D. Daley, "Automatic detection of adhd and asd from expressive behaviour in rgbd data," in *Proc. IEEE International Conference on Automatic Face & Gesture Recognition*, 2017, pp. 762–769.

[42] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, p. 436–444, 2015.

[43] H. Dibeklioğlu, Z. Hammal, and J. F. Cohn, "Dynamic multimodal measurement of depression severity using deep autoencoding," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 2, pp. 525–536, 2018.

[44] M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, and M. Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proc. International Workshop on Audio/Visual Emotion Challenge*, 2016, pp. 3–10.

[45] J. Gratch, R. Artstein, G. Lucas, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, A. Rizzo, and L.-P. Morency, "The distress analysis interview corpus of human and computer interviews," in *Proc. Language Resources and Evaluation Conference*, 2014, pp. 3123–3128.

[46] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "Avec 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proc. International Workshop on Audio/Visual Emotion Challenge*, 2013, p. 3–10.

[47] M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, and M. Pantic, "Avec 2014: 3d dimensional affect and depression recognition challenge," in *Proc. International Workshop on Audio/Visual Emotion Challenge*, 2014, p. 3–10.

[48] D. M. Sloan, M. E. Strauss, S. W. Quirk, and M. Sajatovic, "Subjective and expressive emotional responses in depression," *Journal of Affective Disorders*, vol. 46, no. 2, pp. 135–141, 1997.

[49] A. Z. Brozgold, J. C. Borod, C. C. Martin, L. H. Pick, M. Alpert, and J. Welkowitz, "Social functioning and facial emotional expression in neurological and psychiatric disorders," *Applied Neuropsychology*, vol. 5, no. 1, pp. 15–23, 1998.

[50] J. L. Tsai, N. Pole, R. W. Levenson, and R. F. Muñoz, "The effects of depression on the emotional responses of spanish-speaking latinas," *Cultural Diversity and Ethnic Minority Psychology*, vol. 9, no. 1, p. 49–63, 2003.

[51] T. Hassan, D. Seuß, J. Wollenberg, K. Weitz, M. Kunz, S. Lautenbacher, J.-U. Garbas, and U. Schmid, "Automatic detection of pain from facial expressions: A survey," *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 6, pp. 1815–1831, 2021.

[52] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[53] M. H. Trivedi, "The link between depression and physical symptoms," *Primary Care Companion to the Journal of clinical psychiatry*, vol. 6, pp. 12–16, 2004.

[54] A. Anand, Y. Li, Y. Wang, J. Wu, S. Gao, L. Bukhari, V. P. Mathews, A. Kalnin, and M. J. Lowe, "Activity and connectivity of brain mood regulating circuit in depression: A functional magnetic resonance study," *Biological Psychiatry*, vol. 57, no. 10, pp. 1079–1088, 2005.

[55] T. Deckersbach, D. D. Dougherty, and S. L. Rauch, "Functional imaging of mood and anxiety disorders," *Journal of Neuroimaging*, vol. 16, no. 1, pp. 1–10, 2006.

[56] R. V. Saveanu and C. B. Nemeroff, "Etiology of depression: Genetic and environmental factors," *Psychiatric Clinics of North America*, vol. 35, no. 1, pp. 51–71, 2012.

[57] E. C. Dunn, R. C. Brown, Y. Dai, J. Rosand, N. R. Nugent, A. B. Amstadter, and J. W. Smoller, "Genetic determinants of depression: Recent findings and future directions," *Harvard Review of Psychiatry*, vol. 23, no. 1, pp. 1–18, 2015.

[58] R. C. Kessler, "The effects of stressful life events on depression," *Annual Review of Psychology*, vol. 48, no. 1, pp. 191–214, 1997.

[59] J. A. Russell, "A circumplex model of affect," *Journal of Personality and Social Psychology*, vol. 39, no. 6, pp. 1161–1178, 1980.

[60] W. D. S. Killgore, "Affective valence and arousal in self-rated depression and anxiety," *Perceptual and Motor Skills*, vol. 89, no. 1, pp. 301–304, 1999.

[61] L. A. Clark and D. Watson, "Tripartite model of anxiety and depression: Psychometric evidence and taxonomic implications," *Journal of Abnormal Psychology*, vol. 100, no. 3, pp. 316–336, 1991.

[62] D. Watson and A. Tellegen, "Toward a consensual structure of mood," *Psychological Bulletin*, vol. 98, no. 2, pp. 219–235, 1985.

[63] T. M. Olino, N. L. Lopez-Duran, M. Kovacs, C. J. George, A. L. Gentzler, and D. S. Shaw, "Developmental trajectories of positive and negative affect in children at high and low familial risk for depressive disorder," *Journal of Child Psychology and Psychiatry*, vol. 52, no. 7, pp. 792–799, 2011.

[64] B. F. Chorpita and E. L. Daleiden, "Tripartite dimensions of emotion in a child clinical sample: Measurement strategies and implications for clinical utility," *Journal of Consulting and Clinical Psychology*, vol. 70, no. 5, pp. 1150–1160, 2002.

[65] T. E. Joiner Jr, S. J. Catanzaro, and J. Laurent, "Tripartite structure of positive and negative affect, depression, and anxiety in child and adolescent psychiatric inpatients," *Journal of Abnormal Psychology*, vol. 105, no. 3, pp. 401–409, 1996.

[66] A. T. Beck, R. A. Steer, R. Ball, and W. F. Ranieri, "Comparison of beck depression inventories-ia and-ii in psychiatric outpatients," *Journal of Personality Assessment*, vol. 67, no. 3, pp. 588–597, 1996.

[67] S. El-Den, T. F. Chen, Y.-L. Gan, E. Wong, and C. L. O'Reilly, "The psychometric properties of depression screening tools in primary healthcare settings: A systematic review," *Journal of Affective Disorders*, vol. 225, pp. 503–522, 2018.

[68] M. Dum, J. Pickren, L. C. Sobell, and M. B. Sobell, "Comparing the bdi-ii and the phq-9 with outpatient substance abusers," *Addictive Behaviors*, vol. 33, no. 2, pp. 381–387, 2008.

[69] A. Tadić, I. Helmreich, R. Mergl, M. Hautzinger, R. Kohnen, V. Henkel, and U. Hegerl, "Early improvement is a predictor of treatment outcome in patients with mild major, minor

or subsyndromal depression," *Journal of Affective Disorders*, vol. 120, no. 1, pp. 86–93, 2010.

[70] A. A. Nierenberg, A. H. Farabaugh, J. E. Alpert, J. Gordon, J. J. Worthington, J. F. Rosenbaum, and M. Fava, "Timing of onset of antidepressant response with fluoxetine treatment," *The American Journal of Psychiatry*, vol. 157, pp. 1423–1428, 2000.

[71] M. Vermani, M. Marcus, and M. A. Katzman, "Rates of detection of mood and anxiety disorders in primary care: A descriptive, cross-sectional study," *The Primary Care Companion to CNS Disorders*, vol. 13, no. 2, 2011.

[72] J. Thevenot, M. B. López, and A. Hadid, "A survey on computer vision for assistive medical diagnosis from faces," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 5, pp. 1497–1511, 2018.

[73] Y. Tian, T. Kanade, and J. F. Cohn, "Facial expression recognition," in *Handbook of face recognition*. Springer, 2011, pp. 487–519.

[74] P. Ekman, "Facial expression and emotion," *American Psychologist*, vol. 48, no. 4, pp. 384–392, 1993.

[75] P. Ekman and W. V. Friesen, *Pictures of facial affect*. Consulting Psychologists Press, 1976.

[76] R. E. Jack, O. G. B. Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," *Proceedings of the National Academy of Sciences*, vol. 109, no. 19, pp. 7241–7244, 2012.

[77] D. J. Widlöcher, "Psychomotor retardation: Clinical, theoretical, and psychometric aspects," *Psychiatric Clinics of North America*, vol. 6, no. 1, pp. 27–40, 1983.

[78] J. F. Greden and B. J. Carroll, "Psychomotor function in affective disorders: An overview of new monitoring techniques," *The American Journal of Psychiatry*, vol. 138, no. 11, pp. 1441–1448, 1981.

[79] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, and D. P. Rosenwald, "Social risk and depression: Evidence from manual and automatic facial expression analysis," in *Proc. IEEE International Conference on Automatic Face & Gesture Recognition*, 2013, pp. 1–8.

[80] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[81] M. Tadalagi and A. M. Joshi, "Autodep: Automatic depression detection using facial expressions based on linear binary pattern descriptor," *Medical & Biological Engineering & Computing*, vol. 59, pp. 1339–1354, 2021.

[82] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[83] T. R. Almaev and M. F. Valstar, "Local gabor binary patterns from three orthogonal planes for automatic facial expression recognition," in *Proc. Humaine Association Conference on Affective Computing and Intelligent Interaction*, 2013, pp. 356–361.

[84] R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, and S. Narayanan, "Multimodal prediction of affective dimensions and depression in human-computer interactions," in *Proc. International Workshop on Audio/Visual Emotion Challenge*, 2014, p. 33–40.

[85] A. Dhall and R. Goecke, "A temporally piece-wise fisher vector approach for depression analysis," in *Proc. International Conference on Affective Computing and Intelligent Interaction*, 2015, pp. 255–259.

[86] M. Senoussaoui, M. Sarria-Paja, J. a. F. Santos, and T. H. Falk, "Model fusion for multimodal depression classification and level detection," in *Proc. International Workshop on Audio/Visual Emotion Challenge*, 2014, p. 57–63.

[87] V. Ojansivu and J. Heikkilä, "Blur insensitive texture classification using local phase quantization," in *Proc. International Conference on Image and Signal Processing*, 2008, pp. 236–243.

[88] M. Kächele, M. Glodek, D. Zharkov, S. Meudt, and F. Schwenker, "Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression," in *Proc. International Conference on Pattern Recognition Applications and Methods*, 2014, pp. 671–678.

[89] B. Jiang, M. Valstar, B. Martinez, and M. Pantic, "A dynamic appearance descriptor approach to facial actions temporal modeling," *IEEE Transactions on Cybernetics*, vol. 44, no. 2, pp. 161–174, 2014.

[90] L. Wen, X. Li, G. Guo, and Y. Zhu, "Automated depression diagnosis based on facial dynamic analysis and sparse coding," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 7, pp. 1432–1441, 2015.

[91] L. Yang, D. Jiang, L. He, E. Pei, M. C. Oveneke, and H. Sahli, "Decision tree based depression classification from audio video and language information," in *Proc. International Workshop on Audio/Visual Emotion Challenge*, 2016, p. 89–96.

[92] H. Kaya, F. Çilli, and A. A. Salah, "Ensemble cca for continuous emotion prediction," in *Proc. International Workshop on Audio/Visual Emotion Challenge*, 2014, p. 19–26.

[93] H. Pérez Espinosa, H. J. Escalante, L. Villaseñor Pineda, M. Montes-y Gómez, D. Pinto-Avedaño, and V. Reyez-Meza, "Fusing affective dimensions and audio-visual features from segmented video for depression recognition: Inaoe-buap's participation at avec'14 challenge," in *Proc. International Workshop on Audio/Visual Emotion Challenge*, 2014, p. 49–55.

[94] H. Meng, D. Huang, H. Wang, H. Yang, M. AI-Shuraifi, and Y. Wang, "Depression recognition based on dynamic facial and vocal expression features using partial least square regression," in *Proc. International Workshop on Audio/Visual Emotion Challenge*, 2013, pp. 21–30.

[95] K. Levi and Y. Weiss, "Learning object detection from a small number of examples: The importance of good features," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004, pp. 1–8.

[96] A. Maridaki, A. Pampouchidou, K. Marias, and M. Tsiknakis, "Machine learning techniques for automatic depression assessment," in *Proc. International Conference on Telecommunications and Signal Processing*, 2018, pp. 1–5.

[97] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[98] Y. Zhu, Y. Shang, Z. Shao, and G. Guo, "Automated depression diagnosis based on deep networks to encode facial appearance and dynamics," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 578–584, 2018.

[99] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4724–4733.

[100] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Proc. Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012.

[101] K. M. Kitani, B. D. Ziebart, J. A. Bagnell, and M. Hebert, "Activity forecasting," in *Proc. European Conference on Computer Vision*, 2012, pp. 201–214.

[102] A. Jan, H. Meng, Y. F. B. A. Gaus, and F. Zhang, "Artificial intelligent system for automatic depression level analysis through visual and vocal expressions," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 3, pp. 668–680, 2018.

[103] Y. Kang, X. Jiang, Y. Yin, Y. Shang, and X. Zhou, "Deep transformation learning for depression diagnosis from facial images," in *Proc. Chinese Conference on Biometric Recognition*, 2017, pp. 13–22.

[104] X. Zhou, P. Huang, H. Liu, and S. Niu, "Learning content-adaptive feature pooling for facial depression recognition in videos," *Electronics Letters*, vol. 55, pp. 648–650, 2019.

[105] X. Zhou, K. Jin, Y. Shang, and G. Guo, "Visually interpretable representation learning for depression recognition from facial images," *IEEE Transactions on Affective Computing*, vol. 11, no. 3, pp. 542–552, 2020.

[106] M. Al Jazaery and G. Guo, "Video-based depression level analysis by encoding deep spatiotemporal features," *IEEE Transactions on Affective Computing*, vol. 12, no. 1, pp. 262–268, 2021.

[107] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in *Proc. International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–7.

[108] J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, and D. D. Mehta, "Vocal and facial biomarkers of depression based on motor incoordination and timing," in *Proc. International Workshop on Audio/Visual Emotion Challenge*, 2014, p. 65–72.

[109] G. Littlewort, J. Whitehill, T. Wu, I. Fasel, M. Frank, J. Movellan, and M. Bartlett, "The computer expression recognition toolbox (cert)," in *Proc. IEEE International Conference on Automatic Face & Gesture Recognition*, 2011, pp. 298–305.

[110] S. Song, S. Jaiswal, L. Shen, and M. Valstar, "Spectral representation of behaviour primitives for depression analysis," *IEEE Transactions on Affective Computing*, 2020.

[111] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, "Openface 2.0: Facial behavior analysis toolkit," in *Proc. IEEE International Conference on Automatic Face Gesture Recognition*, 2018, pp. 59–66.

[112] F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, and M. Pantic, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proc. Annual Workshop on Audio/Visual Emotion Challenge*, 2017, p. 3–9.

[113] L. J. Wells, S. M. Gillespie, and P. Rotshtein, "Identification of emotional facial expressions: Effects of expression, intensity, and sex on eye gaze," *Plos One*, vol. 11, no. 12, pp. 1–20, 2016.

[114] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[115] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv:1212.0402*, 2012.

[116] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6546–6555.

[117] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. IEEE International Conference on Computer Vision*, 2019, pp. 6202–6211.

[118] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[119] X. Geng, C. Yin, and Z.-H. Zhou, "Facial age estimation by learning from label distributions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 10, pp. 2401–2412, 2013.

[120] H. Pan, H. Han, S. Shan, and X. Chen, "Mean-variance loss for deep age estimation from a face," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5285–5294.

[121] Y. Zhou, H. Xue, and X. Geng, "Emotion distribution recognition from facial expressions," in *Proc. ACM International Conference on Multimedia*, 2015, pp. 1247–1250.

[122] M. R. Ali, J. Hernandez, E. R. Dorsey, E. Hoque, and D. McDuff, "Spatio-temporal attention and magnification for classification of parkinson's disease from videos collected via the internet," in *Proc. IEEE International Conference on Automatic Face & Gesture Recognition*, 2020, pp. 207–214.

[123] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. British Machine Vision Conference*, 2015, pp. 1–12.

[124] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age," in *Proc. IEEE International Conference on Automatic Face & Gesture Recognition*, 2018, pp. 67–74.

[125] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[126] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[127] L. Liu, Y. Long, P. W. Fieguth, S. Lao, and G. Zhao, "Brint: Binary rotation invariant and noise tolerant texture classification," *IEEE Transactions on Image Processing*, vol. 23, no. 7, pp. 3071–3084, 2014.

[128] K. M. Prkachin and P. E. Solomon, "The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain," *Pain*, vol. 139, no. 2, pp. 267–274, 2008.

[129] S. Kaltwang, O. Rudovic, and M. Pantic, "Continuous pain intensity estimation from facial expressions," in *Proc. International Symposium on Visual Computing*, 2012, pp. 368–377.

[130] C. Florea, L. Florea, and C. Vertan, "Learning pain from emotion: Transferred hot data representation for pain intensity estimation," in *Proc. European Conference on Computer Vision Workshops*, 2014, pp. 778–790.

[131] R. Zhao, Q. Gan, S. Wang, and Q. Ji, "Facial expression intensity estimation using ordinal information," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3466–3474.

[132] J. Zhou, X. Hong, F. Su, and G. Zhao, "Recurrent convolutional neural network regression for continuous pain intensity estimation in video," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 84–92.

[133] J. Yu, T. Kurihara, and S. Zhan, "Frame by frame pain estimation using locally spatial attention learning," in *Proc. Iberian Conference on Pattern Recognition and Image Analysis*, 2019, pp. 229–238.

[134] P. Rodriguez, G. Cucurull, J. Gonzàlez, J. M. Gonfaus, K. Nasrollahi, T. B. Moeslund, and F. X. Roca, "Deep pain: Exploiting long short-term memory networks for facial expression classification," *IEEE Transactions on Cybernetics*, pp. 1–11, 2017.

[135] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[136] M. J.H. Balsters, E. J. Krahmer, M. G.J. Swerts, and A. J.J.M. Vingerhoets, "Verbal and nonverbal correlates for depression: A review," *Current Psychiatry Reviews*, vol. 8, no. 3, pp. 227–234, 2012.

[137] C. Sobin and H. A. Sackeim, "Psychomotor symptoms of depression," *American Journal of Psychiatry*, vol. 154, no. 1, pp. 4–17, 1997.

[138] L. Chao, J. Tao, M. Yang, and Y. Li, "Multi task sequence learning for depression scale prediction from video," in *Proc. International Conference on Affective Computing and Intelligent Interaction*, 2015, pp. 526–531.

[139] M. Senoussaoui, M. Sarria-Paja, J. F. Santos, and T. H. Falk, "Model fusion for multimodal depression classification and level detection," in *Proc. International Workshop on Audio/Visual Emotion Challenge*, 2014, pp. 57–63.

[140] G. Lam, H. Dongyan, and W. Lin, "Context-aware deep learning for multi-modal depression detection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 3946–3950.

[141] A. Ghandeharioun, S. Fedor, L. Sangermano, D. Ionescu, J. Alpert, C. Dale, D. Sontag, and R. Picard, "Objective assessment of depressive symptoms with machine learning and wearable sensors data," in *Proc. International Conference on Affective Computing and Intelligent Interaction*, 2017, pp. 325–332.

[142] A. Seal, R. Bajpai, J. Agnihotri, A. Yazidi, E. Herrera-Viedma, and O. Krejcar, "Deprnet: A deep convolution neural network framework for detecting depression using eeg," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021.

[143] P. Pedrelli, S. Fedor, A. Ghandeharioun, E. Howe, D. F. Ionescu, D. Bhathena, L. B. Fisher, C. Cusin, M. Nyer, A. Yeung, L. Sangermano, D. Mischoulon, J. E. Alpert, and R. W. Picard, "Monitoring changes in depression severity using wearable and mobile sensors," *Frontiers in Psychiatry*, vol. 11, 2020.

[144] M. H. Pollack, "Comorbid anxiety and depression," *Journal of Clinical Psychiatry*, vol. 66, p. 22, 2005.

[145] G. E. Tafet and C. B. Nemeroff, "The links between stress and depression: Psychoneuroendocrinological, genetic, and environmental interactions," *The Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 28, no. 2, pp. 77–88, 2016.

[146] L. J. Iny, J. Pecknold, B. E. Suranyi-Cadotte, B. Bernier, L. Luthe, N. Nair, and M. J. Meaney, "Studies of a neurochemical link between depression, anxiety, and stress from [3h]imipramine and [3h]paroxetine binding on human platelets," *Biological Psychiatry*, vol. 36, no. 5, pp. 251–291, 1994.

[147] Z. Ahmed and S. H. Julius, "The relationship between depression, anxiety and stress among women college students." *Indian Journal of Health & Wellbeing*, vol. 6, no. 12, pp. 1232–1234, 2015.

[148] R. C. Kessler, C. B. Nelson, K. A. McGonagle, J. Liu, M. Swartz, and D. G. Blazer, "Comorbidity of dsm–iii–r major depressive disorder in the general population: Results from the us national comorbidity survey," *The British Journal of Psychiatry*, vol. 168, no. S30, pp. 17–30, 1996.

[149] P. W. Gold, "The organization of the stress system and its dysregulation in depressive illness," *Molecular Psychiatry*, vol. 20, no. 1, pp. 32–47, 2015.

[150] L. Jing and Y. Tian, "Self-supervised visual feature learning with deep neural networks: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 11, pp. 4037–4058, 2020.

[151] M. Tavakolian, M. Bordallo Lopez, and L. Liu, "Self-supervised pain intensity estimation from facial videos via statistical spatiotemporal distillation," *Pattern Recognition Letters*, vol. 140, pp. 26–33, 2020.
[152] X. Yang, X. He, Y. Liang, Y. Yang, S. Zhang, and P. Xie, "Transfer learning or self-supervised learning? a tale of two pretraining paradigms," *arXiv:2007.04234*, 2020.
[153] S. Roy and A. Etemad, *Self-supervised contrastive learning of multi-view facial expressions*, 2021, p. 253–257.
[154] A. Kanazawa, D. W. Jacobs, and M. Chandraker, "Warpnet: Weakly supervised matching for single-view reconstruction," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3253–3261.
[155] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," *arXiv:1803.07728*, 2018.

Original publications

I Carneiro de Melo W, Granger E & Hadid A (2019) Combining global and local convolutional 3D networks for detecting depression from facial expressions. Proc. IEEE International Conference on Automatic Face & Gesture Recognition (FG), Lille, France, pp. 1–8.

II Carneiro de Melo W, Granger E & Hadid A (2019) Depression detection based on deep distribution learning. Proc. IEEE International Conference on Image Processing (ICIP), Taipei, Taiwan, pp. 4544–4548.

III Carneiro de Melo W, Granger E & Hadid A (2020) A deep multiscale spatiotemporal network for assessing depression from facial dynamics. IEEE Transactions on Affective Computing, doi:10.1109/TAFFC.2020.3021755.

IV Carneiro de Melo W, Granger E & Bordallo Lopez M (2020) Encoding temporal information for automatic depression recognition from facial analysis. Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Barcelona, Spain, pp. 1080–1084.

V Carneiro de Melo W, Granger E & Bordallo Lopez M (2021) MDN: A deep maximization-differentiation network for spatio-temporal depression detection. IEEE Transactions on Affective Computing, doi: 10.1109/TAFFC.2021.3072579.

VI Carneiro de Melo W, Granger E & Bordallo Lopez M (2022) Automatic facial expression analysis using decomposed multiscale spatiotemporal networks. Submitted for evaluation. ArXiv preprint arXiv:2203.11111, https://doi.org/10.48550/arXiv.2203.11111.

Original publications are not included in the electronic version of the dissertation.

820. Nissilä, Tuukka (2022) Ice-templated cellulose nanofiber structures as reinforcement material in composites

821. Yu, Zitong (2022) Physiological signals measurement and spoofing detection from face video

822. Chen, Haoyu (2022) Human gesture and micro-gesture analysis : datasets, methods, and applications

823. Khan, Iqra Sadaf (2022) Exploring Industry 4.0 and its impact on sustainability and collaborative innovation

824. Peng, Wei (2022) Automatic neural network learning for human behavior understanding

825. Zhang, Ruichi (2022) Vanadium removal and recovery from liquid waste streams

826. Väätäjä, Maria (2022) Prospects of the room temperature fabrication method for electroceramics : feasibility for printing techniques and integration with temperature-sensitive materials

827. Li, Yante (2022) Machine learning for perceiving facial micro-expression

828. Behzad, Muzammil (2022) Deep learning methods for analyzing vision-based emotion recognition from 3D/4D facial point clouds

829. Leppänen, Tero (2022) From industrial side streams to the circular economy business : value chain, business ecosystem and productisation approach

830. Tuomela, Anne (2022) Enhancing the safety and surveillance of tailings storage facilities in cold climates

831. Kumar, Dileep (2022) Latency and reliability aware radio resource allocation for multi-antenna systems

832. Niu, He (2022) Valorization of mining wastes in alkali-activated materials

833. Jayasinghe, Laddu Praneeth Roshan (2022) Coordinated multiantenna interference mitigation techniques for flexible TDD systems

834. Zhu, Ruixue (2022) Interaction peculiarities of red blood cells and hemorheological alterations induced by laser radiation

835. Kuosmanen, Elina (2022) Technological support for Parkinson's disease patients' self-care

836. Li, Jing (2022) Advanced high and low field ^{1}H and ^{129}Xe NMR methods for studying polymerization, curing and pore structures of geopolymers

Book orders:
Virtual book store
http://verkkokauppa.juvenesprint.fi

UNIVERSITY OF OULU P.O. Box 8000 FI-90014 UNIVERSITY OF OULU FINLAND

ACTA UNIVERSITATIS OULUENSIS

UNIVERSITY OF OULU

ISBN 978-952-62-3366-6 (Paperback)
ISBN 978-952-62-3367-3 (PDF)
ISSN 0355-3213 (Print)
ISSN 1796-2226 (Online)