Pattern Analysis and Applications
https://doi.org/10.1007/s10044-021-00988-8

ORIGINAL ARTICLE

Deep transfer learning in human–robot interaction for cognitive and physical rehabilitation purposes

Chaudhary Muhammad Aqdus Ilyas[1,2] · Matthias Rehm[2] · Kamal Nasrollahi[1,4] · Yeganeh Madadi[1,3] · Thomas B. Moeslund[1] · Vahid Seydi[3]

Received: 11 July 2020 / Accepted: 30 April 2021

Abstract

This paper presents the extraction of the emotional signals from traumatic brain-injured (TBI) patients through the analysis of facial features and implementation of the effective emotion-recognition model through the Pepper robot to assist in the rehabilitation process. The identification of emotional cues from TBI patients is very challenging due to unique and diverse psychological, physiological, and behavioral challenges such as non-cooperation, facial/body paralysis, upper or lower limb impairments, cognitive, motor, and hearing skills inhibition. It is essential to read subtle changes in the emotional cues of TBI patients for effective communication and the development of affect-based systems. To analyze the variations of the emotional signal in TBI patients, a new database is collected in a natural and unconstrained environment from eleven residents of a neurological center in three different modalities, RGB, thermal and depth in three specified scenarios performing physical, cognitive and social communication rehabilitation activities. Due to the lack of labeled data, a deep transfer learning method is applied to efficiently classify emotions. The emotion classification model is tested through closed-field study and installment of a Pepper robot equipped with the trained model. Our deep trained and fine-tuned emotional recognition model composed of CNN-LSTM has improved the performance by 1.47% on MMI, and 4.96% on FER2013 validation data set. In addition, use of temporal information and transfer learning techniques to overcome TBI-data limitations has increased the performance efficacy on challenging dataset of neurologically impaired people. Findings that emerged from the study illustrate the noticeable effectiveness of SoftBank Pepper robot equipped with deep trained emotion recognition model in developing rehabilitation strategies by monitoring the TBI patient's emotions. This research article presents the technical solution for real therapeutic robot interaction to rehabilitate patients with standard monitoring, assessment, and feedback in the neuro centers.

Keywords Deep transfer learning · Emotion recognition · Traumatic brain injury (TBI) · TBI patients database · Cognitive, social, and physical therapy · Rehabilitation strategies · Human–robot interaction · Assistive care · Assessment and monitoring · Augmentative and Assistive Technology (AAT)

1 Introduction

It is challenging for people with traumatic brain injury (TBI) to communicate and socialize due to motor, hearing, and speech inhibitions. For rehabilitation and training purposes, TBI-patients are often treated in specialized neuro centers. Since 2015, our researchers have been working with a national neuro center with a focus on providing technical systems enhancing capability for the residents and to provide assistance and facilitation to staff members [27, 28, 64]. The majority of the residents at the neuro center possess unique and highly diverse nature of impaired cognitive and behavioral abilities (for instance, apraxia and aphasia). As

✉ Chaudhary Muhammad Aqdus Ilyas
cmai@create.aau.dk

[1] Visual Analysis and Perception (VAP) Lab, Aalborg University, Aalborg, Denmark

[2] Human–Robot Interaction (HRI) Lab, Aalborg University, Aalborg, Denmark

[3] Department of Computer Engineering, Faculty of Technical and Engineering, South Tehran Branch, Islamic Azad University, Tehran, Iran

[4] Research Department of Milestone Systems A/S, Copenhagen, Denmark

Published online: 22 June 2021

some of these residents are unable to recover from their life-altering impairments fully, the center provides full-time care and aid in organizing and supporting activities of daily living (ADL). Providing such facilities is resource, labor, and expertise expensive. It also produces extra strain on the staff members to maintain the same level and standard of services to these residents. One technical means of lifting this burden is intelligent augmented and assistive technologies (AAT) that can be of help to maintain the quality of services and to facilitate staff members in developing and implementing rehabilitation strategies.

Researchers focus on providing assistance in naturals environments through ambient assisted living (AAL). AAL contributes in wide utility space such as from patients to social services, health workers to smart homes and multi-agent systems with the aim to present a solution for independent living in the user's preferred living environment [12]. AAL aims to provide better quality of life for both elderly people and their care-workers. Recent advances such as the adoption of Internet-of-Things (IoT), cloud computing (CC), virtual and augmented reality (VAR), ambient intelligence (AmI) and neurorobotics have tackled the AAL solutions. According to [48], IoT technologies in the AAL domain are capable of catering to challenges related to ADL, elderly care, social dis-cohesion, personalized medication, physical activities, health tracking and various other applications. In addition, brain computer interface (BCI) systems contribute to improve the quality of life of elderly people by receiving and transmitting brain signals to external aids and VAR devices [3]. However, the major limitation of employing BCI systems involves wearable sensors mounted on the head to communicate signals to the linked devices, which restricts natural movement of the subjects under observation.

In the AAL domain, researchers have developed specialized AAT systems tailor-made for completion and facilitation of specific tasks such as robots for surgical-operations [10, 75], healthcare robots for monitoring elderly people [63], social assistive robots for social engagement, e.g., for children with autism spectrum disorders (ASD) [7, 11, 66], or human–computer interfaces for assistance in daily tasks [63]. The AAT systems, specifically developed for elderly care or disabled people, employ different input signals to process information like audio, video, proximity, touch, and their combination is based upon the system application and environment. Over the past few decades, researchers are exerting special efforts to develop such systems with more human-like characteristics like social assistive robots (SAR), to assist in ADL. SARs can be integrated with emotional signal recognition and synthesis for natural and human-like interaction.

There are various ways to extract emotional signals, as one of the regions of the brain stem cell (amygdala) is mainly responsible for generating actions related to emotional arousal [1]. We can identify the activation of signals through this brain region by reactions visible through external and internal body stimuli. For instance, the amygdala regulates the release of hormones in the bloodstream, controls the heart rate, blood pressure, skin conductance, as well as changes in facial expressions [54]. In a nutshell, we can determine these emotional cues by dilation of eye-pupil, electron flow on the skin (skin conductance), brain activity (electroencephalography (EEG)), magnetic resonance imaging (MRI), heart rate (electrocardiography (ECG)), and facial expression recognition (FER) [4] as demonstrated in Fig. 1. Many researchers focus on the various techniques for the rehabilitation of physical and cognitive impaired people, e.g., [67] establish a virtual reality exposure therapy (VRET) for managing stress reactions. Similarly, [37] develop a BCI system for the extraction of psychological signals of mentally impaired people using electroencephalography (EEG). For developing an affect-based system for use in rehabilitation settings, the real challenge thus lies in the acquisition of emotional signals from people suffering from neurological disorders like patients with acquired brain injury.

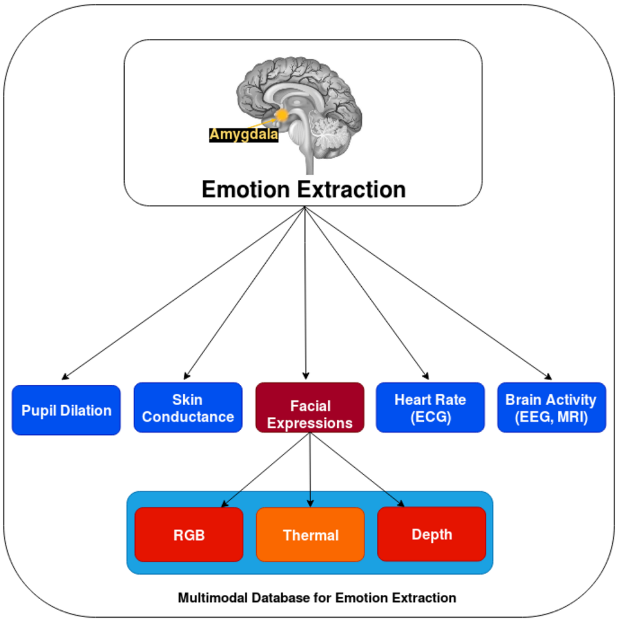


Fig. 1 Emotional signal identification through various parameters; collection of data through multi-modal channels to analyze facial expressions

Considering the challenges associated with this user group such as limited muscle movement or paralysis, non-cooperative behavior, inappropriate responses, impaired reasoning, involuntary head, and upper body movements, mental inflexibility with non-compliance, agitation, loud verbalization and sometimes physical aggression, we decided to collect emotional signals in an unobtrusive manner through facial expression analysis [27–29]. Other methods to identify emotional signals have certain limitations like the installment of sensors on the body, e.g., for identifying emotions through skin conductance, ECG, MRI, and EEG. Pupil dilation measurements involve an eye-tracking camera that must be placed close to the face without any occlusion, which is not possible due to limitations related to the physiology of the residents.

Therefore, considering the challenges mentioned above and complexities associated with TBI patients and aiming at capturing data in the natural environment, we extracted emotional signals through facial expressions relying on Ekman's definition of basic emotions. Ekman et al. described six basic expressions (anger, disgust, fear, happiness, sadness, and surprise) as universal basic emotional cues among humans [19]. Details of the data acquisition system in the specified scenarios, modified strategies for improved data quality and pre-processing techniques are mentioned in Sect. 3

The automatic recognition of facial expressions and interpretation as emotional cues can be utilized in a broad spectrum of socially and emotionally sensitive systems such as robots and virtual humans that engage with people in real-world contexts naturally. Since real-world frameworks encompass uncontrolled settings, where businesses operate in continuous altering circumstances such as occlusions, noise, illumination variations, diverging facial postures, and unwanted head and body movements. Therefore, systems that execute automatic analysis of human emotions must be robust to visual-data acquisition conditions, varying contexts, and the time of response.

In the past few decades, the performance of automatic-facial expression recognition (A-FER) systems was limited to controlled conditions and posed expressions. These systems were exploiting facial information that is captured in laboratory environment with majority of induced expressions such as Cohn-Kanade database [76], Cohn-Kanade extended database [46], MMI database [79], JAFFE database [47], DISFA database [50], and DISFA extended database [49] and therefore less prone to environmental challenges like illumination and occlusion, pose variation, and spontaneous expressions. Recently researchers are exerting extra effort to develop systems that could perform A-FER in natural circumstances. For this purposes, scientists have collected database in-the-wild such as AFEW [15], SFEW [17], FER2013[23], ExpW [91], and BU-3DFE and BU-4DFE databases[84, 90]. In addition, emotion recognition challenges are carried out to address the challenges in real-world scenarios. However, researchers have illustrated that facial expression (FE) in naturalistic interaction thoroughly varies from the induced or posed ones [14, 29, 68, 87]. Additionally, facial expressions of the TBI patients have additional artifacts such as facial paralysis, non-cooperation during data acquisitions, and large pose variation [28]. Therefore, these databases have the following limitations:

- The databases collected under controlled environmental conditions, with proper illumination and cooperative subject, contain frontal postures in the majority of images or minimal pose variation (Figs. 2, 3). However, acquiring data from real-life patients, suffering from brain injuries, is remarkably complex as patients are not cooperative, and it is quite difficult to have frontal postures. Moreover, facial databases captured in-the-wild have diverse features as compared to database captured in the laboratory environment. Therefore, FER systems trained under "controlled conditions" do not perform well in real-world applications. So it is essential to build a database of TBI patients in natural and unconstrained circumstances.
- Facial expressions of TBI patients significantly vary as compared to healthy people due to prolonged disabilities, paralysis, and continued state of depression. Researches

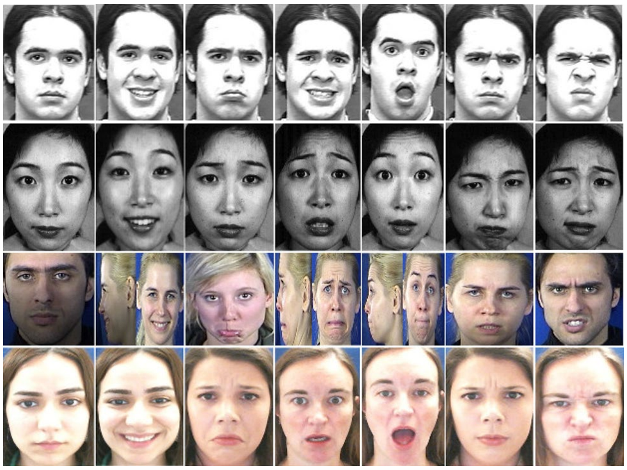


Fig. 2 Sample database images captured in controlled conditions for facial expressions: Databases (rows top to bottom) CK+ , JAFFE, MMI and DISFA+ ; Emotion categories; (from left to right) Neutral, Happy, Sad, Fear, Surprise, Angry & Disgust (For CK+ contempt)

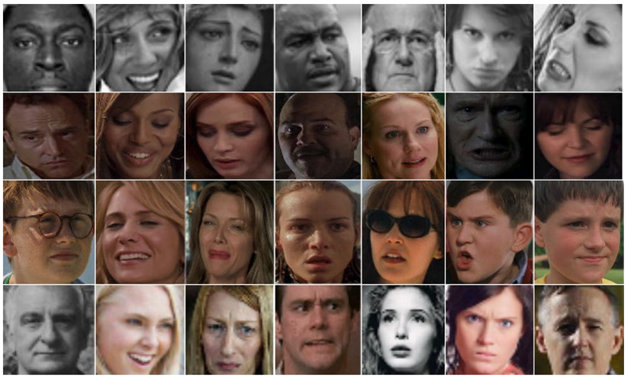


Fig. 3 Sample database images captured in uncontrolled conditions for facial expressions: Databases (rows top to bottom) FER2013, AFEW, SFEW, ExpW ; Emotion categories; (from left to right) Neutral, Happy, Sad, Fear, Surprise, Angry & Disgust

have associated the dominance of negative expressions with this user group [71]. Furthermore, existing FER databases have induced expression, that is different from natural expressions produced involuntary [14, 29, 68, 87].
- Some of the TBI patients have additional complexities due to facial paralysis, so their expressions are quite hard to extract. In addition, some of the facial-symmetry and facial bones of the TBI patients are misaligned due to the stroke. Images with such features are not available in current databases.
- Facial expressions of healthy people are easily distinguishable such as happiness, sadness, anger, fear, surprise, disgust, and neutral. TBI patients do not have clear six expressions, but we find a prominence of only two to three expressions, usually the negative ones. It is essential for deploying affect-based intelligent interactive systems with these users that systems are trained on a specially dedicated database, developed in real environmental conditions with all the complexities associated due to the brain injury and real-world challenges.

In this paper, we aim to address the limitations mentioned above by the development of a TBI patient database under natural, unconstrained, and uncontrolled conditions. This multimodal visual database is collected with RGB, thermal, and depth sensors in the specific scenarios to ensure uniformity and reliability in data collection. Database annotation is performed by the neuro center staff members, experts, caregivers, physiotherapists, and doctors, who worked with a particular resident for more than six months. It contains a range of expressions from the residents performing daily activities like physiotherapy, cognitive rehabilitation activities, and social communication. We have collected 1723 videos in 91 sessions, illustrating emotional reactions of 11 subjects in three modalities: RGB, thermal, and depth.

There exists a vast range of emotional and facial expression recognition databases. However, they have limitations, mostly because data are acquired in controlled laboratory environments. Additionally, all of the existing databases are of healthy people with quite clear expressions that are remarkably different from brain-injured residents of the neuro center, who do not show the same variation in the six basic expressions. To reach more realistic and exact results, we developed the TBI patient database. As we know, learning deep NNs need massive labeled training data. So we applied a deep transfer learning model to utilize related data from other databases to help the training the model.

The main contributions of the paper are as follows:

- This research article focuses on the extraction of psychological signals of neurologically impaired people using transfer learning (TL) techniques that assist the care-workers to monitor and assess the rehabilitation process with increased emotional efficacy.
- The research article contributes to designing a specialized framework for collecting consistent and reliable data from neurologically impaired people for social, physical, and cognitive well-being.
- We employed a deep architecture of CNN and CNN plus RNN to develop a FER model. This FER model is tested on CK+, MMI, JAFFE, FER-2013, AFEW, SFEW2.0, DISFA, and ExpW databases and competes with the state-of-the-art methods and outperforms some of them.
- It is demonstrated that the deep trained FER model is capable of recognizing emotions of people with facial paralysis in a natural environment, producing state-of-the-art performances.
- Integrating the FER model with the SoftBank Pepper robot to recognize emotions helps the staff members and care workers to understand the emotional conditions of the residents better and adopt the rehabilitation and interaction strategies in real-time.
- Our findings indicate that the robot intervention with the residents of the neuro center enhanced the productivity of physiotherapy and social interaction.

The rest of the paper is organized as follows: Section 2 provides an overview of existing databases and related research in the field of facial expression recognition (FER) with the focus on natural data collection environment. Section 3 explains the process of data collection of brain-injured patients in various scenarios. Section 4 presents the methodologies implemented in our approach. Section 5 describes the experimental studies and result evaluation. Section 6 illustrates the contribution toward rehabilitation strategies. Section 7 concludes the paper.

2 Related work

2.1 Current databases

Existing databases of facial expression recognition such as Cohn-Kanade (CK, CK+) [46, 76], MMI [79], CE [18], JAFFE [47], and BU-4DFE [84, 90] are developed in laboratory and controlled conditions where subjects displayed distinctive facial expressions. These databases have high-quality-based posed-expressions. However, non-posed and spontaneous expressions acquired in uncontrolled or in-the-wild environments are quite different from posed expressions. It is essential to identify non-posed expressions in a natural or uncontrolled environment for automatic affective computing. Thus, researchers focused toward data acquisition in-the-wild or uncontrolled settings such as AFEW and SFEW datasets [17], used in series of EmotiW challenges[1], or FER-2013 [23], DISFA [50], DISFA+ [49]. These

[1] https://sites.google.com/site/emotiwchallenge/

databases encompass multimodal effects such as voice, biological parameters, and sequences of frames. However, due to the number of subjects, pose variation, and environmental settings, the range of diversification of these databases is minimal. We briefly describe the databases that are captured in-the-wild as well as in controlled settings (Tables 1, 2), used for emotion recognition, and will discuss their limits leading to the creation of the TBI database.

CK+ database The extended Cohn-Kanade (CK+) database [46] is one of the most extensively used databases for FER systems. It is established in the laboratory or controlled settings, with 593 image sequences of 123 subjects, of which only 327 are annotated with seven emotion labels (six basic emotions and contempt). The database consists of 69% females and 31% males with an age range from 18 to 50 years. The dataset contains posed and non-posed facial expressions at a maximum intensity level.

Table 1 An overview of the facial expression databases

Databases	No. of Sub.	Samples	Env.	Nature (posed/spontaneous)	Expressions information	Availability
CK+ [46]	123	593 image sequences	Controlled (laboratory)	Posed and spontaneous	6 Basic expressions (with contempt) plus neutral	http://www.consortium.ri.cmu.edu/ckagree/
JAFFE [47]	10	213 Images	Controlled (laboratory)	Posed	6 Basic expressions plus neutral	https://zenodo.org/jaffe
MMI [79]	25	740 Images 2900 videos	Controlled (laboratory)	Posed	6 Basic expressions plus neutral	https://mmifacedb.eu/
DISFA [50]	27	89,000 images	Controlled (laboratory)	Spontaneous	AU-FACS (6 Basic expressions plus neutral (by EMFACS system))	http://mohammadmahoor.com/disfa
FER2013 [23]	N/A	35,887 images	Web (in-the-wild)	Posed and spontaneous	6 Basic expressions plus neutral	https://www.kaggle.com/fer2013
AFEW [15]	330	1809 videos	Movies (in-the-wild)	Posed and spontaneous	6 Basic expressions plus neutral	https://sites.google.com/view/emotiw2018/home
SFEW2.0 [16]	N/A	1766 images	Movies (in-the-wild)	Posed and spontaneous	6 Basic expressions plus neutral	https://cs.anu.edu.au/few/AFEW.html
ExpW [91]	N/A	91,793 images	Web (in-the-wild)	Posed and spontaneous	6 Basic expressions plus neutral	http://mmlab.ie.cuhk.edu.hk/projects/socialrelation/index.html

Table 2 Number of data images for each expression for the databases

Database	CK+	JAFFE	MMI	DISFA	AFEW2018		FER2013		SFEW		ExpW	
Image size	640 * 490 720 * 480	256 * 256	720 * 576	768 * 1024	N/A		48 * 48		720*576		N/A	
F-Exps					Training	Val	Training	Val	Training	Val	Training	Val
Anger	90	30	1959	436	118	59	4953	958	178	77	1272	318
Contempt	36	0	0	0	0	0	0	0	0	0	0	0
Disgust	0	29	1517	5326	72	39	547	111	66	23	1250	312
Fear	50	32	1313	4073	76	44	5121	1024	98	47	329	82
Happy	138	31	2785	28,404	142	63	8989	1774	198	73	10,576	2644
Neutral	324	30	3034	48,582	129	61	6198	1233	150	86	8309	2077
Sad	56	31	2169	1024	104	59	6077	1247	172	73	2494	623
Surprise	166	30	1746	1365	70	46	4002	831	96	57	2471	617
Total	860	213	14,523	89,210	711	371	35,887	7178	958	436	26,701	6673

MMI database The MMI database [79] is captured in the laboratory or controlled settings with 326 image sequences of 32 subjects. Two hundred thirteen image sequences are labeled with six basic expressions with onset-apex-offset states.

JAFFE The Japanese Female Facial Expressions (JAFFE) [47] database is captured in controlled conditions. It consists of 213 image samples of 10 female subjects. Each subject has 3–4 facial images with each of six basic expressions and one image with a neutral expression.

DISFA Denver Intensity of Spontaneous Facial Actions (DISFA) database [50] consists of 27 subjects captured with spontaneous expressions. It is coded with Action Units (AUs) ranges from 0 to 5 with zero corresponding to the absence of any activation of muscles, while five belongs to maximum intensities. We have employed the EMFACS conversion system [22] to convert AU FACS codes to emotional expressions that presented approximately 89,000 images with a majority having neutral expressions.

EmotiW-AFEW-2018 Acted Facial Expressions in the Wild (AFEW) [17] and its subset Static Facial Expressions in the Wild (SFEW) [16] have been used as a benchmark dataset for annual emotion recognition in the wild challenge (EmotiW) challenge. AFEW is a multimodal-temporal database containing facial expressions from movies and reality TV shows that are close to real-world scenarios. AFEW consists of 330 subjects with an age range from one to seventy-seven years (1–77 yrs). The annotation of this database is according to six basic expressions (anger, disgust, fear, happiness, sadness, and surprise) and a neutral expression. The AFEW 7.0 dataset used in EmotiW 2017 consists of subject independent data partitions with training (773 samples), validation (383 samples) and test sets (653 samples).

SFEW Static Facial Expressions in the Wild (SFEW) [16] is developed by extracting few images from AFEW with varied head poses, close to real-life illumination conditions, age-range, and distinctive facial expressions. The SEFW 2.0 is used in the EmotiW 2015 challenge, and it is most commonly used in general. The dataset is divided into three partitions: training set (958 image samples), validation set (436 image samples), and test set (372 image samples). Each image sample is assigned with one of seven basic expressions, i.e., anger, disgust, fear, happy, neutral, sadness, and surprise.

FER-2013 The FER-2013 database [23] consists of approximately 36,000 images, labeled with seven emotion classes (six Ekman emotional states plus neutral expression). The database is established by using Google image search combined with phrases for gender, age, ethnicity, and 184 emotion-related keywords. FER-2013 is one of the biggest databases for FER in-the-wild environment but with a low image resolution of 48 * 48 pixels leading to problems for facial landmark detectors.

EXPW The Expression in-the-wild (ExpW) database [91] is comprised of approximately 90,000 facial images downloaded from the web. Each of the images is manually assigned to one of the seven primary expressions.

Nonetheless, all the databases, as mentioned earlier, consist of images of healthy people without any facial paralysis, cognitive or physiological impairments (Figs. 2, 3). Hence, there is a need for the development of systems dedicated to cognitive and physical impaired persons like TBI patients, based on natural, spontaneous, unposed, and uninduced facial expressions. To address these demands, we developed a database of TBI residents in natural and uncontrolled settings, details provided in Sect. 3.

2.2 Current architectures for affect recognition

Automatic affective computing is a well-established research area, and there are a wide variety of algorithms and databases to develop automated affect recognition mechanisms. We would like to briefly discuss state-of-the-art methods for emotion-related search on the databases explained in Sect. 2.1. Emotion recognition systems can be distinguished by the methods employed for feature extraction and feature classification. Most of the advanced FER systems are exploiting the techniques of convolutional neural networks (CNN) for facial feature extraction and classification (Table 3), as they provide state-of-the-art results for facial expression recognition [6, 9, 44], pain identification [6] and interpretation as emotional states [13, 80]. Conventional algorithms for affect recognition use handcrafted features such as pixel intensities [53], Gabor filters [8], local binary patterns (LBP) [69, 92], local quantized pattern (LQP) [78] and histogram of oriented gradients (HoG) [2]. Handcrafted features are accompanied by unintended features that have no or less impact on classification. In the case of handcrafted features, not all possible cases can be included for features selection and classification, so its performance is compromised.

The significant advantage of deep learning methods over conventional machine learning models is the simultaneous performance of feature extraction and classification. Moreover, deep learning methods apply iterative approaches for feature extraction and optimize error by back propagation, thus resulting in those critical features that human experts can miss while handcrafting the features. Recently used deep learning algorithms for FE and emotional analysis have demonstrated a remarkable ability to learn features and achieved state-of-the-art results in a range of learning tasks like cross-database evaluation where handcrafted features exhibit low performances due to lack of generalization to new scenarios. Moreover, deep neural networks perform remarkably well for subject independent estimation schemes of emotional expression recognition. This interdependence contributes to

Table 3 Summary of architectures and methods for affect recognition

Method	Database	Architecture
Mohammadi et al. [53]	CK+, MMI	Sparse representation classification, PCA-based dictionary building
Shan et al. [69]	MMI, JAFFE	Boosted local binary pattern (B-LBP) + SVM
Zhao and Zhang [92]	CK, JAFFE	Kernel discriminant isometric mapping (KDIsomap)
Liu et al. [43]	EmotiW-2014 (AFEW)	Multiple Riemannian kernels + SVM
Liu et al. [44]	CK+, JAFFE	Boosted deep belief networks (BDBN)
Yao et al. [82]	EmotiW-2015 (audio–video) SFEW, AFEW	Emotional expression relation and facial muscle activation unit (AU) with RBF kernel
Kaya et al. [33]	EmotiW-2015 (audio–video) AFEW, AFEW	Partial least squares regression (PLS) and kernel extreme learning machines (ELM) with multi-level weighted fusion
Ng et al. [55]	EmotiW-2015	Transfer learning for deep CNN, pre-trained on the ImageNet dataset; cascading fine-tuning
Yao et al. [83]	Emotiw-2016	HoloNet, CNN with concatenated rectified linear unit (CReLU)
Rodriguez et al. [65]	CK+	VGG-16 + LSTM
Yan et al. [81]	AFEW6.0, CHEAVD (audio–video)	Multi-cue fusion; cascaded CNN and Bi-directional-RNN CNN + SVM
Liu et al. [45]	CK+, MMI, SFEW	CNN with loss layers
Li et al. [42]	CK+, SFEW,	Deep locality-preserving CNN (DLP-CNN)
Zhang et al. [91]	CK+, SFEW, FER-2013	CNN with multi-task network (MN)
Kim et al. [34]	FER-2013	Discriminative deep CNN (DCNNs); alignment-mapping networks (AMNs); CNN with network ensemble
Meng et al. [52]	CK+, MMI, SFEW	CNN with MN; identity-aware CNN (IACNN)
Yu and Zhang [85]	SFEW	CNN with network ensemble
Zhao et al. [93]	CK+	Expression intensity-invariant network (EIN)
Yu et al. [86]	CK+	Expression intensity-invariant network (EIN) + multi-task-CNN (MN)
Kim et al. [35]	CK+, MMI	Expression intensity-invariant network (EIN) with data augmentation, illumination normalization and face frontalization
Zhang et al. [89]	CK+, MMI,	Network ensemble with cascaded CNN and SDM
Kuo et al. [39]	CK+	Applied FA network and Intraface
Sun et al. [72]	MMI	NE with GoogLeNet and SDM
Otberdout et al. [57]	AFEW	Deep CNN + symmetric positive definite (SPD) matrices
Fan et al. [21]	AFEW	CNN with VGG-LSTM and fusion techs

the formulation of this paper, as the stability and reliability of the deep learning systems could perfectly align with the procedures required for clarifying complexity in emotion analysis in natural and unconstrained environments, mainly dealing with brain-injured patients.

Deep neural networks, notably CNNs, are well-established approaches for researchers in the field of deep-vision for FER. In the FER-2013 challenge, [74] achieved the 1st prize by exploiting deep neural networks in two stages: use of CNN trained in a supervised way at a first stage and a second stage applying support vector machines (SVM) on the output of the trained CNN. Kahou et al. [31] winner of the EmotiW-2013 challenge, used the CNN and deep belief network (DBN) composed of two-stacked layers of restricted Boltzmann machines (RBMs). The first layer of RBM comprised Gaussian RBM with noisy ReLU, and the second layer Gaussian-Bernoulli RBM. This method worked well and managed to get the at-the-time state-of-the-art performance but at higher computation cost for larger datasets. In 2014, [44] incorporated three tasks of feature learning, feature selection, and classification in a unified manner by employing Boosted Deep Belief Networks (BDBN) and managed to achieve remarkable results in challenging conditions. The winner of the EmotiW-2014 challenge [43] combined multiple kernels on Riemannian manifolds for emotion classification by the measurement of corresponding similarities and distances. Researchers in [43] employed SVM, logistic regression, and least-squares models for emotion classification and applied decision level fusion. However, along with high computation cost for feature extraction, this method produced lower accuracy when exposed to challenging emotional categories.

Kulkarni et al. [38] demonstrated the good results to determine whether 6-class expressions are genuine or these facial movements are fake. He addresses the problem by projecting facial features in deeply learnt space. However, 12 class and the binary emotion pair classification problem still remains a challenge. This is because the distinguishing factors between the unfelt and genuine expressions occur in a very short part of the whole emotion and are a challenge to model. Guo et al. [24] presented dataset with 50 classes of compound emotions for affective computing and geometrically represented the landmark displacement to recognize emotions. However, it is challenging to determine dominant or complementary emotions. Yao et al. [82] explored the significance of the suppressed relationship between evolving characteristics derived from facial muscle motions. The particular relations and patterns between emotional expression and facial muscle activation unit (AU) are extracted and called it AU-Aware facial features. This method leads them to surpass the EmotiW-2015 challenge without using additional data. [33] applied two least-squares regressions, specifically partial least square (PLS) and Kernel extreme learning machines (ELM) with multi-level weighted fusion for emotional classification. One of the drawbacks of applying multi-level fusion with different input modalities audio or video could result in performance downgrading. [55] applied transfer learning techniques on a small dataset for static facial expression recognition in the wild, by pre-training their network on ImageNet dataset followed by fine-tuning to target dataset and achieved comparable results.

In the year 2016, [83] applied a deep but computational efficient CNN with concatenated rectified linear unit (CReLU) and inception-residual structural for emotional recognition under unconstrained environment. In the year 2017, [65] exercised CNN to learn features from VGG-Faces and integrated with long short-term memory (LSTM) to gain the temporal information. This approach was further improved by [6], who applied deep CNN for features classification into expressions and fed the system with super-resolved facial images. [81] employed the cascaded CNN and RNN, where images are first fed into CNN for facial features extraction, followed by bidirectional RNN to learn the changes. One of the common aspects in the work of the [6, 55, 65, 83] the use of extensive annotated data of healthy people, captured in controlled and uncontrolled environmental conditions. Transfer learning can be applied to overcome the challenges of training CNNs that require large annotated training datasets of diverse expressions. Transfer learning overcomes the limited data problem by transferring image features learned with CNNs on large datasets to other visual recognition tasks on targeted, limited training data samples [56]. In the case of TBI database, transfer learning is applied to learn features from large-scale public datasets captured in varied environmental conditions and distinct scenarios, with the presence of all expression states, to serve as a better weight initialization by fine-tuning.

The work in [55, 56, 65, 81] exhibited state-of-the-art results for emotional challenges, but healthy subjects. Therefore, we investigated a similar approach for the TBI dataset. We employed CNN pre-trained to VGG-16 to learn the features from eight public databases and then by applying transfer learning approaches, fined-tuned to TBI dataset to overcome the identity and unbalanced emotional-data limitations.

3 Traumatic brain injured people database (TBI-database)

3.1 Data acquisition

Data were collected at a neuro center that offers 24/7 rehabilitative care for their residents with brain injury. The goal was to record visual data from the residents in natural scenarios to extract emotional information. Due to the nature of their impairments, it is very complex to collect data for all expressions of anger, sadness, happiness, surprise, and disgust. Moreover, residents have diverse cognitive, physical, and interactive skills. Sometimes the residents demonstrate physical and verbal aggression along with inappropriate responses. Most of the computer vision techniques for FER are dependent on data quality and environmental conditions like occlusion, lighting, and face pose and alignment. Considering these conditions, we collected the data in three different scenarios with the help of experts, trainers, and caregivers to have reliable and the best possible quality of the data in unconstrained scenarios. These situations are (a) cognitive rehabilitation strategies, (b) physical rehabilitation strategies, and (c) social interaction aiding strategies. Generally, a caregiver follows a set of protocols [5] for the rehabilitation tasks.

In order to deploy automated affect-based systems based on facial expressions, it is necessary to set up a signal perceiving sensors-system, in our design RGB, thermal, and depth sensors. However, there is no extensive research explaining data collection methods for the FE of people that have suffered from TBI residents.

The studies in [58, 79] explained database creation and organization of healthy and cooperative subjects with spontaneous and induced expressions in a controlled laboratory environment or in-the-wild settings or through online websites. However, in the case of our residents, there is no database, or database development protocols, so we relied on data acquisition with rehabilitation protocols and then modified them after analyzing them carefully. We set up the data acquisition system with RGB, thermal, and depth cameras, placed at 1.5 meters distance from the residents

Table 4 Subjects in database along with challenges due to TBI, number of sessions and activities

Subjects	No. of sessions	Activities			Challenges			Prominent features
		Cognitive	Physio	Social	Body paralysis	Speech inhibition	Facial paralysis	
A	12	4	4	4	Complete	Yes	Partial	High anger
B	10	4	3	3	Left side	No	No	High arousal
C	10	4	3	3	Lower body	No	No	Excessive head movement
D	9	3	3	3	Partial	No	Partial	Emotionally unstable
E	9	2	4	3	No	Yes	Partial	Emotionally unstable
F	7	2	3	2	Partial	No	No	High arousal
G	6	2	2	2	Lower body	No	No	Excessive upper body movement
H	7	2	3	2	No	No	Partial	Low arousal
I	6	2	2	2	Yes	Yes	Partial	Low arousal
J	8	2	3	3	No	No	No	Verbal and physical aggression
K	7	3	3	1	Partial	Yes	No	Emotionally unstable

while performing their rehabilitation and social activities. Experts prescribe playing games as a therapy is the most effective way to aid brain injury recovery [20, 60, 70]. Researchers recommend five games for brain injury recovery: Card games, Sudoku, Lumosity, TherAppy, and Tetris [61]. We modified these games, including other rehabilitation activities to obtain optimal data for the training of a deep learning-based system; details are provided in the later Sects. 3.1.1–3.1.3

Data collection approaches are distinguished by the rehabilitation activities and the disability of the resident. We collected data from eleven residents. The precise nature of their disability is described in Table 4. Due to severe and diverse conditions of these residents with emotional instability, experts plan strategies for their recovery based on their health conditions and neuropsychological test results [5, 77]. Furthermore, these residents have impaired facial and emotional expressions, accompanied by frequent mood swings, low concentration (Table 4), and significant pose variations in regards to the capture of facial images.

It is also challenging to extract all six basic expressions, so to have useful facial video data, we altered the standard rehabilitation activities to gather more diverse information.

3.1.1 Cognitive rehabilitation strategy

The basic aim of this activity is to improve the ability of residents to understand and interpret information to perform specific functions mentally. Emotional stability is a key factor in this training; otherwise, residents will not be able to participate and get the advantage of these exercises. For this purpose, caregivers follow a set of protocols like Mini-Mental State Exam (MMSE)[2] and Montreal Cognitive Assessment (MoCA)[3] comprised of repetitive activities with gradual increase in difficulty level, to assess the attention, memory [62], visuospatial perception [51], language and communication, function execution and learning ability of brain-injured residents [77]. These tasks are mostly accomplished through the use of calendars, drawing clocks, memory log or memory devices, alarms or reminders, reading or listening to books, and playing games. The majority of these activities were performed on the paper placed on a table. During these activities, we encountered a couple of problems that resulted in poor data quality: a) subjects mostly looked downwards, b) frequent pose changes, and c) less attention. Hence, these rehabilitation tasks were tailored to the requirements of the residents in the following ways:

- Residents performed the tasks on a PC tablet, as mentioned earlier, that was placed in parallel to the cameras, which resulted in more frontal facial images and increased attention.
- A favorite movie clip or cartoon character of a resident was displayed on the screen, and then residents were asked about the character or the story. This activity was repeated, and the cognitive assessment was monitored accordingly.
- Error-less (EL) learning was performed by instructing residents to sing lyrics of songs, match pictures, stack Lego bricks, and play computer games, which are of the subjects' interest.

[2] https://www.sundhed.dk/sundhedsfaglig/laegehaandbogen/undersoegelser-og-proever/skemaer/geriatri/mms-mini-mental-status/

[3] https://www.mocatest.org/

- Sudoku is an organizational game with numbers, colors or alphabets, normally played on paper. Residents played this game electronically on the tablets placed at a predefined location and orientation, resulting in frontal facial images. Most of the residents found the game apparatus comfortable, and there was a wide range of games from easy to hard difficulty providing the opportunity for trainers to monitor the learning skills of the resident at each level.
- Older residents preferred card games rather than playing digitally. Therefore, card games like Memory, Solitaire, Go-fish, and war were played with them. These games proved to be beneficial in recovery as they involve strategy and thought processes with smaller challenges [61]. Regularly playing these games boosted memory skills as well as mathematical understanding, depending on the game. Cognitive skills assessors confirmed this result.
- We have introduced another application based game 'Lumosity' for improved memory, problem-solving, and to speed-up processing. This app presents the range of brain training games based on the input information to improve learning skills. Residents showed a positive response to this app.
- Residents suffering from speech problems were asked to play TherAppy, an application based game developed by Tactus Therapy Solutions, created for residents' language skills recovery. This game comprises of four modules for Comprehension, Naming, Reading, and Writing [73]. Residents were asked to recall the name of a picture, complete a phrase, or spell a word after listening to a short sound clip. Hints were available by clicking a button if a resident was struggling.
- Most of the residents exhibit negative expressions like sadness, depression, anger, or aggression more frequently. In order to have other expressions like surprise, happiness, or joy, various games were created in such a way that intentionally lead to winning for the residents that resulted in more positive expressions.

Attention and memory enhancement are core elements in mental training. All these modified strategies were implemented on eleven residents, generated less erroneous database, and the residents exhibited more expressions and learning as compared to the custom exercises for cognitive skills recovery. Cognitive skills were evaluated by meeting goals and levels of mental-games applications. Performance evaluation is discussed in detail in Sect. 5

3.1.2 Physical rehabilitation strategy

TBI causes physical morbidity due to damage to the sensory-motor system. Depending on the nature of the damage, it can cause reduced muscle movement and paralysis to the upper limb, lower limb, or complete body. Physical rehabilitation methods are planned case to case while considering age, gender, disability type, and post-concussion symptoms [26]. Additionally, assessment of activity tolerance (Table 4), balance, coordination, and postural control estimation are taken into account while conducting cardiovascular, muscular-skeletal, and vestibular activities. Physiotherapists conduct these activities through preset operations like cardio exercises, using a treadmill, walking or mild running independently or with a trainer, cycling, push-ups, squats, and other related exercises after assessing the abilities of residents [26]. During all these activities, facial data are hardly available due to the excessive movement of the body or face. Therefore, to acquire the maximum facial data, we asked residents who do not have or have partial paralysis to perform physical exercises:

- Residents ride a stationary bicycle to have a static upper body as much as possible while looking at a tablet placed parallel to cameras. During the exercise, expressions were recorded.
- For residents who use wheelchairs, the tasks were designed accordingly, so they moved their chair forward and backward within three meters for multiple sessions.
- Activities such as hand press-ups, arm raises, and cup pick-up and placing were performed.
- Console video games were also introduced, which aided the movement of the resident arms and hands to a certain extent while playing. These games exhibited more explicit expressions and hand-eye coordination.
- Card games also helped with training dexterity and gross motor skills.

These activities resulted in useful data while enhancing the interest of residents throughout the therapy sessions.

3.1.3 Social rehabilitation strategy

Social rehabilitation is quite a complex and long-term challenge due to cognitive and behavioral disorders. Social reintegration strategies are based on individual cognitive progress, mental health, and behavioral distortions. In a standard scenario at the neuro center, the residents sit around a table over a cup of tea and share their daily activities. Often, residents do not take an interest, and trainers have to intervene by asking questions. Another observed problem is that residents with speech inhibition communicate through writing letters on tablets, which slows down communication and reduces interest. To overcome these challenges, we introduced the following activities:

- Firstly, we shared storybooks with the residents and asked them to read aloud to other residents of the neuro center. Most of the participants did not take an interest in listening to the story due to poor storytelling skills and limited concentration.

- Secondly, we played card games with residents resulting in better interaction with the other participants as compared to the storytelling activity.
- Thirdly, we utilized PS4 console games. Every participant showed interest individually or as part of a team. Most of the participants enjoyed Medal of Honor Airborne (MOHA)[4], Need for Speed[5] and similar games. When playing MOHA in two teams, participants of each team worked closely with each other, enhancing mutual interaction. They also expressed their emotions better at the different stages of the games.

These activities also helped in physiotherapy. However, it is still challenging to get all the emotional states due to non-cooperation, traumatic disabilities, and other social and technical issues; therefore, we have further classified the expressions into positive and negative expressions [28].

Data are collected in multiple phases throughout 91 sessions, as presented in Table 4 with RGB, thermal, and depth sensors. In total, we collected 1723 video events, each of a maximum of 5 s in length.

3.2 Data annotation

Furthermore, for accurate annotations, only those experts or trainers were consulted who worked with these residents for more than three months and have at least ten months of experience dealing with residents that suffered from brain injury. Experts annotated the videos manually and then later verified when image sequences are split into various categories. Various pre-processing steps are applied to develop a high-quality facial database; details are provided in Sect. 4.

4 Methodology

In this section, we describe the three main steps for the automatic recognition of facial expressions (FE), i.e., pre-processing, facial feature learning, and facial features classification. The algorithms explored and state-of-the-art implementations for processes, as mentioned earlier, are presented below:

4.1 Pre-processing

Pre-processing is a vital step to avoid unwanted features for facial expression recognition, such as illumination variations, background clutter, and different head poses. Therefore, to ensure the learning of only essential features, we applied the following pre-processing algorithms before exposure to neural networks training for the formation of a high-quality facial data log.

4.1.1 Face-alignment

The first step for FER tasks is face detection to remove background and non-relevant features. Viola-Jones (VJ) [30] is a classical method, widely used for face detection that is robust and accurate for frontal faces. However, the algorithm exhibits lower performance in natural and in-the-wild environments, where faces are not always frontal, producing false detection. To achieve higher quality data, we have used the dlib-CNN-face detector [36] that has surpassed VJ for face detection, under unconstrained and natural environmental conditions with significant pose variations [88]. In addition, for further face alignment, we have estimated the facial landmarks through a cascaded regression method, i.e., supervised descent method (SDM), which tracks 49 facial points and reduces the variations and in-plane rotation.

4.1.2 Illumination and pose normalization

Deep neural networks are sensitive to illumination and contrast, which can lead to significant intra-class variations even when the images of the same person displaying the same expressions have different contrast and illumination. We have employed histogram equalization combined with illumination normalization, as this method has produced state-of-the-art results in the literature of FER [41]. Another challenge, associated with unconstrained and natural settings, are facial images with large pose variations. We have employed the pose normalization technique that produces frontal views, where landmarks are calculated with arbitrary facial positions, and by finding the inverse of the transpose matrix, the face is frontalized [25].

4.2 Deep learning architecture for feature learning and transfer learning (convolutional neural network)

Our work is focused on the emotion cues from images and sequences of images. Convolutional layers are richly embedded with spatial information. We have used the features from convolution layers instead of fully connected layers and transferred to the target database for fine-tuning. To take advantage of temporal information, we have utilized the long short-term memory (LSTM) network to consider the sequences of CNN actuations explicitly. CNNs like VGG-16 and AlexNet, which are pre-trained on ImageNet, can be used as a feature extractor.

[4] https://www.ea.com/games/medal-of-honor

[5] https://www.ea.com/games/need-for-speed

Spatial feature extraction In order to make full use of static databases, we have used VGG-16 architecture for dimensional feature extractions. The VGG-16 is the deep convolutional network with up to sixteen layers (thirteen convolutional layers and three fully connected layers). This network takes an input image size of 224 * 224 pixels, with a convolutional kernel size of 3 * 3 and max-pooling with 2 * 2 windows. We used the pre-trained VGG-Face [59] architecture to initialize the network parameters that are trained on a massive facial dataset of 2.6 million images. We assume that databases that are captured in controlled and uncontrolled environmental conditions with posed as well as spontaneous expression are involved, and we use the transfer learning strategy to transfer the "information" learned by the VGG-model to our new target dataset of neuro center residents suffering from brain injuries for emotional cues identification. Transfer learning can be used to avoid overfitting in the training of our network (Fig. 4), considering the TBI database is too limited in terms of identities of subject to train a generalized network.

The LSTM for temporal information extraction In general, CNNs deal with images that are isolated. However, in our case, we have used sequences of images as well, thus preserving the temporal information. LSTM models are capable of absorbing this dynamic sequential information. The LSTM modules can determine long-range temporal correlations from the input sequences by using memory cells, which can hold and release information.

As illustrated in Fig. 5, the LSTM states are controlled by three gates associated with forget (*f*), input (*i*), and output (*o*) states. These gates regulate the flow of information through the model by using point-wise multiplications and sigmoid functions σ, which bind the information flow between zero and one by the set of mathematical equations as explained in [27, 28].

The datasets used to train the CNN were chosen from the benchmark datasets publicly available or made available to the research community, and they are described in Sect. 2.1.

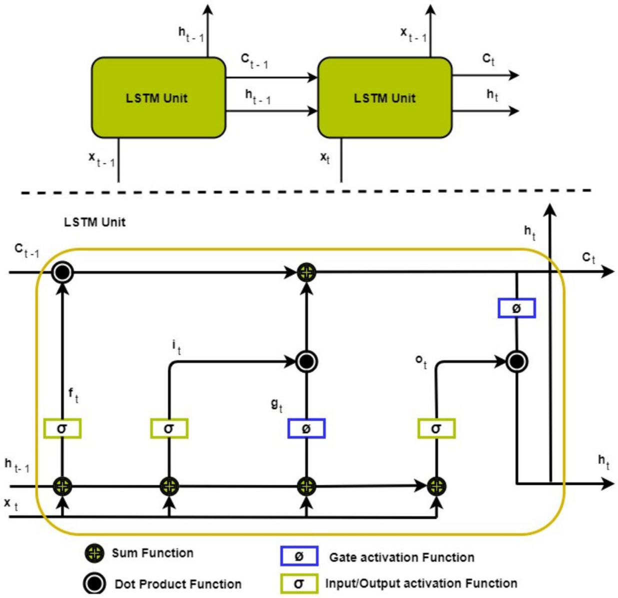


Fig. 5 LSTM architecture with memory unit

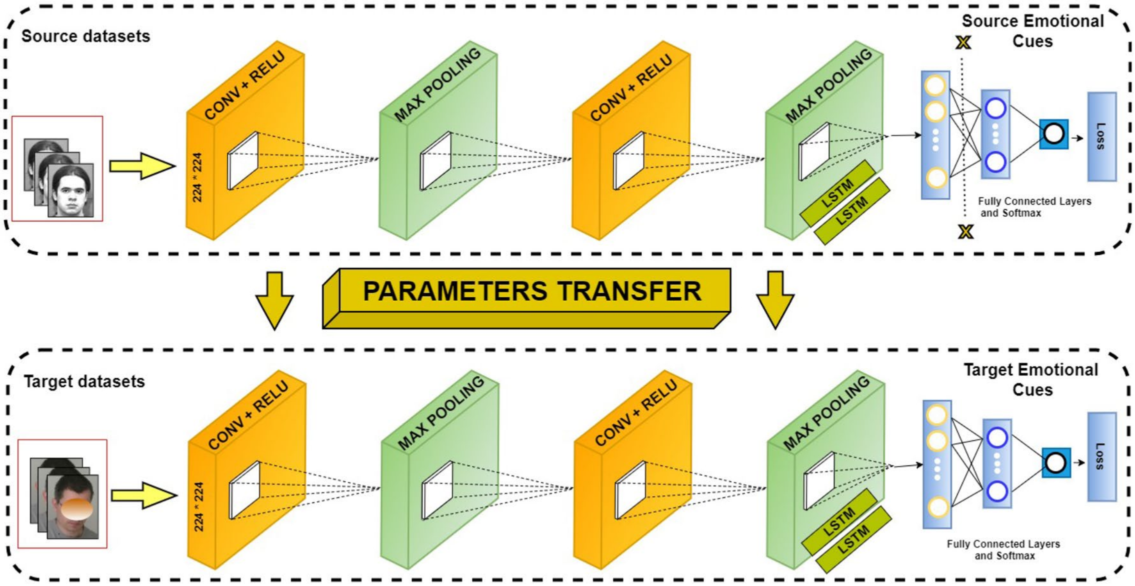


Fig. 4 Transfer learning model architecture

4.3 Transfer learning mechanism

In the current research project, we have to deal with limited labeled and identity data from people that suffered a traumatic brain injury. However, learning processes in deep neural networks need lots of labeled training data. Gathering training data and labeling it is very difficult and time-consuming work. So, for gaining more accurate results, we make use of new techniques such as transfer learning.

Transfer learning is a powerful technique which adapts knowledge from some related auxiliary well-labeled source domains. Considering the benefit of transfer learning, we can use labeled data that was gathered with healthy subjects to optimize target data. In general, transfer learning methods categorize into two groups: domain-invariant feature learning and classifier adaptation. In this paper, we applied an in-depth transfer learning approach to unify the knowledge transfer and deep feature learning.

Since the input of our architecture is image frames and image sequences, we had implemented the learning of features in two ways: firstly by the use of only static images and transferring the knowledge to the TBI datasets; secondly exploiting the dynamic features of video data, as the variations between image sequences encode additional advantageous information for the classification of emotional signals.

Similar to the work in [55, 65, 81], we employed the VGG-16 model to initialize the network parameters and learn the features from eight public databases. Since the bottom layers of CNNs learn more generic features and top layers acquire more sophisticated and data specific information[32], we reserved only the convolutional and max-pooling layers and discarded the pre-trained last three fully connected layers. We removed fully connected layers as they do not hold spatial information (Fig. 6), which is essential for the capture of motion signals in the subsequent LSTM model. Therefore, the last pooling layers of the CNN framework are linked directly to the LSTM to investigate the temporal characteristics across coherent images.

5 Experimental results

In this section, we evaluate the performance of our proposed model in two ways: first, by the domain transfer learning of static as well as dynamic databases to our target TBI database; second, by evaluating the emotional cues learned and transferred from controlled and uncontrolled environmental conditions to the TBI datasets. A static dataset refers to the image frames, whereas dynamic relates to the sequences of images or video sequences.

5.1 Experimental results evaluation for static datasets

The facial images are resized to 224 * 244 pixels according to the network-input parameter. Peak expression frame is used for training of the network for CK+, MMI, DISFA+ datasets. JAFFE, FER2013, SFEW, ExpW have mostly one to four images per expression. Video datasets are first converted into 30 frames per second by an open-source video

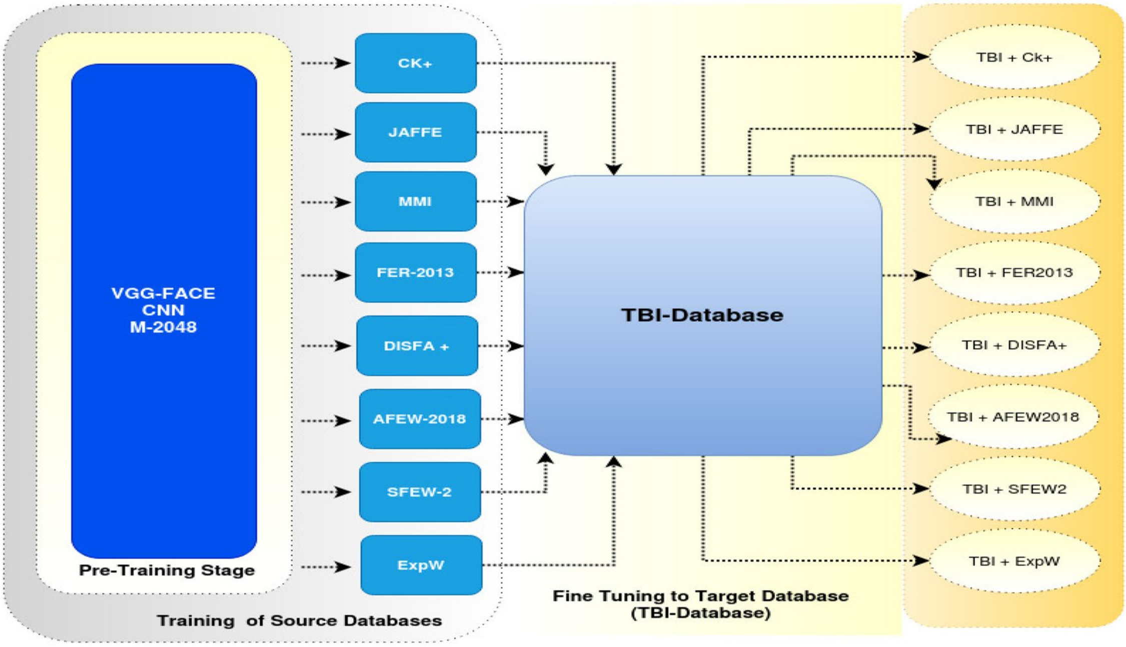
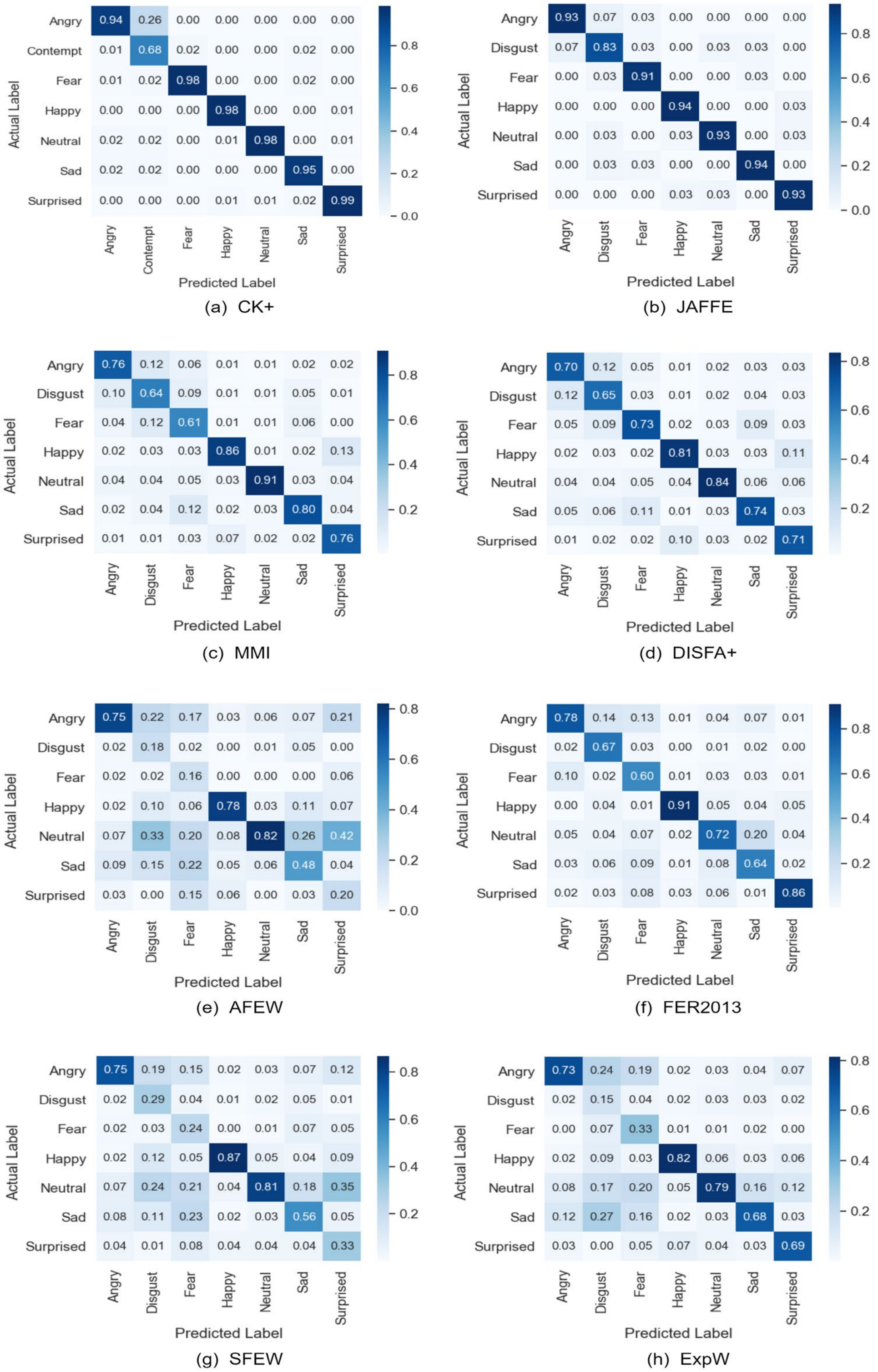


Fig. 6 Databases explored for transfer learning

converter, and then the peak expression image is selected. Data are distributed 80% for training and 20% for testing purposes. The network is trained with a learning rate of 0.0001, and batch normalization is applied to normalize the input layer.

Figure 7 illustrates the performance of our models trained on eight different datasets. We can identify that recognition performance of contempt is not good as compared to other expressions through the confusion matrix in Fig. 7a. Besides, we can determine that fear and disgust emotion expressions are less accurate, as demonstrated by the confusion matrix in Fig. 7c. However confusion matrix of datasets captured in controlled environment Fig. 7a–d have much higher performance than of in-the-wild setting databases as evident in



(a) CK+

(b) JAFFE

(c) MMI

(d) DISFA+

(e) AFEW

(f) FER2013

(g) SFEW

(h) ExpW

Fig. 7 Performance visualization of the models trained on eight source databases using image frames

Fig. 7e–h for emotional categories. The overall accuracies of our proposed network are compared with other state-of-the-art methods, as seen in Table 5, and it is observed that our model has performed competitively well.

5.2 Dynamic database

The temporal information exploration is analyzed on four publicly available datasets, namely CK+, MMI, DISFA+, and AFEW. The performance of fine-tuned VGG-face model is compared with state-of-the-art methods in Table 6. It is clearly observed that in the case of the DISFA+ dataset, our network has produced better results. Similarly, our network has surpassed the state-of-the-art methods in case of AFEW dataset, when tested on the validation set. For CK+ and MMI datasets, our fine-tuned model produced decent and competitive results. The confusion matrices to represent the accuracies of seven emotional categories are illustrated in Fig. 8. Figure 9 represents the performance of our architecture employed to static and dynamic datasets. It is evident that temporal information has increased the performance of the network.

Table 5 Performance evaluation of our (VGG-FineTuned) model for emotional categories for static datasets with other results in the literature in terms of average accuracy

Group	Method	Training Parameters	Accuracy (%)
CK+	Liu et al. [45]	Eight folds	97.1
	Zhang et al. [91]	Ten folds	**98.9**
	Our	Ten folds	98.6 ± 0.59
JAFFE	Liu et al. [44]	LOSO	**91.8**
	Our	Ten folds	89.46 ± 1.75
MMI	Liu et al. [45]	Ten folds	78.53
	Li et al. [42]	Five folds	78.46
	Our	Ten folds	**79**.06 ± **0**.88
DISFA+	Our	Five folds	**77**.15 ± **4**.92
FER 2013	Zhang et al. [91]	Training 28,709 Validation 3589 Test 3589	Test 75.1
	Tang [74]		Test 71.2
	Kim et al. [34]		Test 73.73
	Our	Training 35,887 Validation 7178	Val **78**.19 ± **2**.47
SFEW	Li et al. [42]	Training 958, Validation 436, Test 372	Val 54.19 (47.97)
	Meng et al. [52]		Val 50.98 (42.57) Test: 54.30 (44.77)
	Yu and Zhang [85]		**Val 55.96 (47.31)** **Test 61.29 (51.27)**
	Our		Val 55.75 ± 2.74

Bold values highlight the maximum accuracy achieved by a certain method on a specific dataset. We have also highlighted our results to show they have achieved either state-of-the-art performance or competed well with other state-of-art methods

5.3 Contribution in emotion recognition

In the second stage, the target 'TBI datasets' are fine-tuned with pre-trained and tuned VGG-face model with the above-mentioned publicly available datasets, in both static and dynamic formats. Despite the challenges of less-expressing and limited-identity datasets, fine-tuned model exhibited the comparable results. In our experimentation, we executed single-source-single-target transfer learning, that is individual source dataset features are transferred to TBI dataset and then emotions are classified. Our network learned the facial features related to the specific emotional category of healthy people and explored those characteristics into facial features of TBI-datasets.

5.4 Evaluation metric

We evaluated the performance of our framework using evaluation matrices to fully understand the model efficacy. Confusion matrices, precision, recall, Area under curve (AUC), and the average accuracy present the performance of our model to recognize subtle emotional changes. We calculated multi-class confusion matrices for both static and dynamic datasets as well as before and after fine tuning to the target datasets as shown in Figs. 7, 8, 11, and 12.

To understand the strength of each dataset for a particular expression category, we employed precision and recall matrices as illustrated in Table 7 and Table 8. Results demonstrate that dataset captured in the wild such as AFEW, SFEW, and ExpW have lesser accuracy for disgust, fear and surprise expressions. However, FER2013 performed quite well for the same expressions. We identified that mis-classification of these emotions could be due to a lower number of such expressions in the datasets under analysis. A trend of increase in accuracy for each emotional class is witnessed with an increase in number of frames. Overall, the precision-recall matrices work in relationship; precision indicate the ability of model to determine only relevant data points whereas recalls verify that determined data points are actually relevant.

As given in the equations, we determined accuracy, precision and recall:

$$\text{Accuracy} = \frac{\text{TP} + \text{TF}}{\text{TP} + \text{TF} + \text{FP} + \text{FN}} \tag{1}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

Table 6 Performance evaluation of our (VGG-finetuned) model for emotional categories for dynamic datasets with other results in the literature in terms of average accuracy

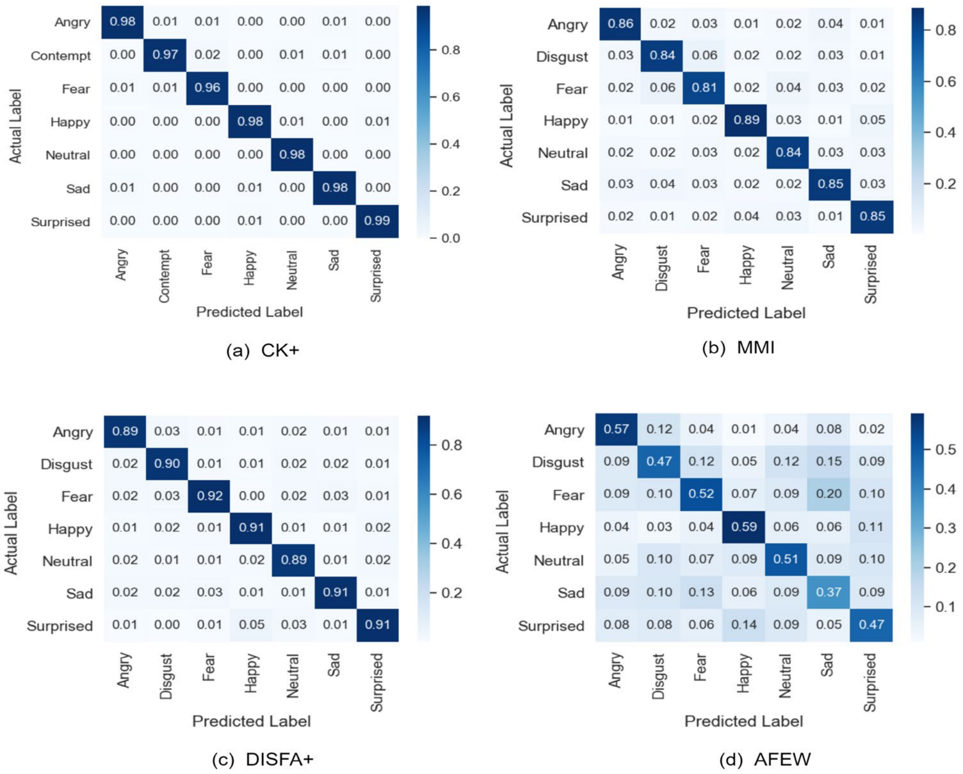
Group	Method	Training Parameters	Accuracy (%)
CK+	Zhao et al. [93]	Training: 7 to last frame Test: last frame; Ten folds	99.3
	Yu et al. [86]	Training: 7 to last frames Test: peak expression; Ten folds	**99.6**
	Kim et al. [35]	All frames used in training and testing; Ten folds	97.93
	Zhang et al. [89]	All frames used in training and testing; Ten folds	98.50
	Kuo et al. [39]	9 frames for training and testing; Ten folds	98.47
	Our	Ten folds	98.92 ± 0.32
MMI	Kim et al. [35]	LOSO	81.53
	Zhang et al. [89]	All frames for training and testing; Ten folds	81.18
	Sun et al. [72]	Ten folds	**91.46**
	Our	Ten folds	85.89 ± 1.52
DISFA+	Zhang et al. [89]	All frames for training and testing; Ten folds	93%
	Our	Ten folds	**94**.09 ± **0**.77
AFEW	Otberdout et al. [57]	Training 773, Validation 373, Test 593 videos	Val 46.32 Test 49.59
	Fan et al. [21]		45.43 on Val
	Fan et al. [21]		59.02 on test
	Our		Val **50**.17 ± **1**.68

Bold values highlight the maximum accuracy achieved by a certain method on a specific dataset. We have also highlighted our results to show they have achieved either state-of-the-art performance or competed well with other state-of-art methods

where TP, TN, FP, and FN are the overall true positive, true negative, false positive, and false negative of all the classes in the confusion matrix. In other words, the overall accuracy was the sum of off-diagonal elements divided by all the elements in the multi-class confusion matrix. Table 5 demonstrates the performance of our model on static source datasets. It is evident that our model has performed better with accuracy of 79.06% and 78.19%, surpassing 78.53% and 73.73% on MMI and FER2013-validation dataset, respectively. On the rest of datasets, our model competed state-of-art-methods while measuring frame-based accuracies. On contrast, our model with sequential information has fared well surpassing recognition accuracies by 94.09% and 50.17% on the DISFA and the AFEW datasets respectively as presented in Table 6. We have used average accuracy metric due to imbalanced emotional data as mentioned in Table 1. For additional performance measure, statistical significance of emotional recognition is verified by t-test conducted on all datasets.

Figure 9 provides the illustration of overall performance of the network exploiting the static as well as temporal information from the various source datasets. It also demonstrates the performance of the network on the target TBI challenging dataset after fine-tuning with source datasets. It is evident from the results that use of temporal information have enhanced the accuracy as it is evidenced through AUC metrics in Fig. 10, where static and temporal information are considered in the model training. In addition, fine-tuning with various source datasets exhibited that performance is dependent on two factors: One is more training data facilitates better in transfer of features and secondly, features related to negative emotions are learnt better from the datasets captured in controlled settings. It is seen from the confusion matrices in Figs. 11 and 12, that accuracy of emotional expressions of anger, contempt/disgust and fear is better when fine-tuned with CK+, MMI, and DISFA+ as compared to AFEW.

Fig. 8 Performance visualization of models trained on four source databases using sequences of images



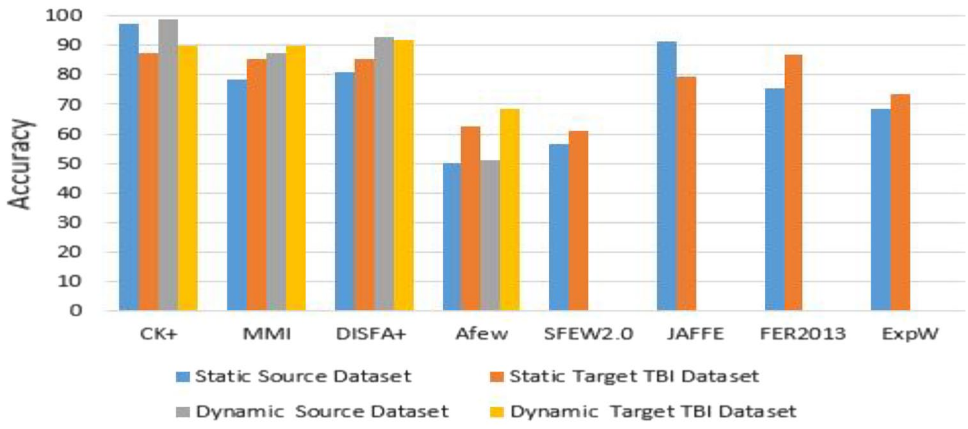
(a) CK+

(b) MMI

(c) DISFA+

(d) AFEW

Fig. 9 Source versus target datasets accuracy comparison: Illustration provides the performance of the network when static and temporal information from both source and target datasets is utilized. For CK+, MMI, DISFA+ and AFEW datasets we have used both static and dynamic information, whereas for SFEW, JAFFE, FER2013 and ExpW static information is explored



6 Insights on emotion recognition in the rehabilitation of TBI patients

The rehabilitation phase usually requires four steps [40]. First, the impairment type and its severity must be tested. Second, the therapist set rehabilitation goals. Third, the rehabilitation intervention takes place. Finally, following the intervention, the patient has to be re-evaluated, allowing to adjust the objectives. Robots have the potential to assist and promote rehabilitation procedures. They can be used to measure performance prior, during and after an intervention as well as systematically and continuously suggest treatment strategies based on this input and the severity of the disability. The intervention of the Pepper robot integrated with customized emotion recognition module assisted the rehabilitation process for the TBI patients in the four phases, as mentioned earlier. In our field study, first, we studied the impairment severity of each patient, and pre-set targets were defined and tested during and after the intervention of the pepper robot. In our case, we distribute the pepper robot assistance in two categories; robot as a monitoring agent and as a feedback agent for both patients and therapists.

Table 7 Precision matrix for each expression class for source datasets

	Precision							
Expressions	CK+	JAFFE	MMI	DISFA	AFEW	FER2013	SFEW	ExpW
Anger	96.67	90.01	78.87	69.72	77.55	79.31	77.78	74.49
Contempt	88.89	–	–	–	–	–	–	–
Disgust	–	82.76	67.56	64.814	18.5	68.68	29.59	66.03
Fear	96	93.75	66.79	72.91	16.32	61.45	24.51	33.67
Happy	98.56	96.77	83.95	81.03	83.83	92.89	88.64	83
Neutral	98.76	90	85.13	83.54	83.67	73.77	82.89	81.25
Sad	94.64	93.55	77.82	73.63	48.45	64.57	57.60	71.38
Surprise	96.75	93.32	77.42	71.42	20.40	87.60	32.85	70.47

Table 8 Recall matrix for each expression class for source datasets

	Recall							
Expressions	CK+	JAFFE	MMI	DISFA	AFEW	FER2013	SFEW	ExpW
Anger	94.56	93.10	75.95	71.53	50.67	66.95	57.48	62.39
Contempt	86.48	–	–	–	–	–	–	–
Disgust	–	82.66	63.86	66.96	66.67	90.66	70.31	82.5
Fear	94.56	88.91	61.24	5574	64.21	7.64	61.53	84.22
Happy	97.84	93.75	85.76	92.23	70.33	83.63	72.13	79.04
Neutral	98.43	93.01	90.98	95.68	37.61	65.17	43.38	55.03
Sad	96.36	91.31	79.92	23.32	44.34	70.01	50.96	60
Surprise	98.77	93.33	75.98	17.74	48.75	82.07	59.25	79.57

6.1 Pepper robot as a monitoring agent

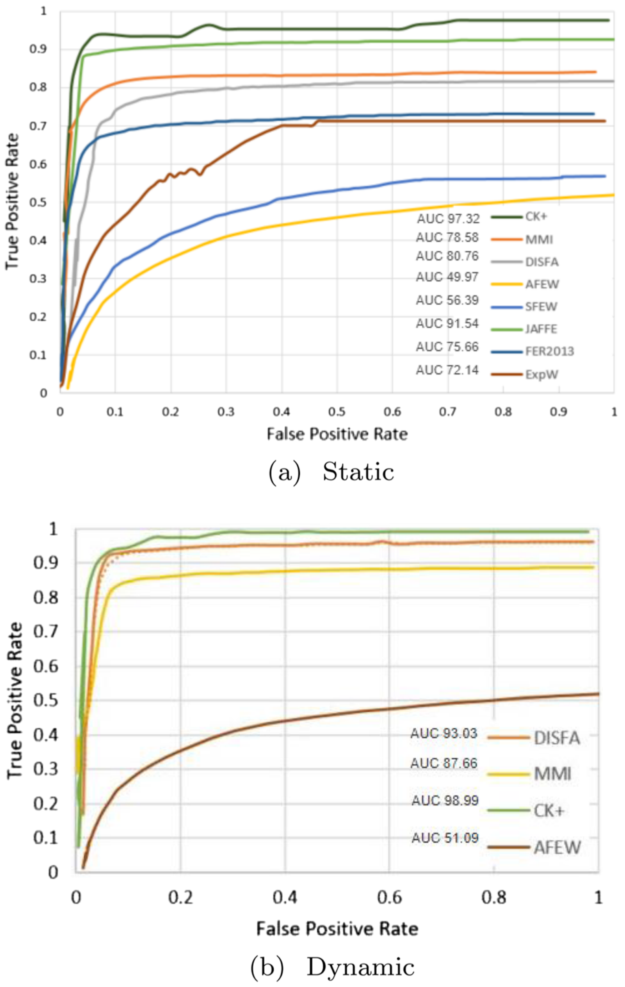
The intervention with the Pepper robot has been designed to in relation to the three scenarios used during the data collection: cognitive, physical and social interaction rehabilitation. The first phase involves the determination of the impairment level for each scenario. It is determined with the set of protocols and disability condition as mentioned in the table 4. In our pilot study we determine the emotional expressions before, during and after each rehabilitation strategy. Before the deployment of the pepper interventions, the data collected was extremely beneficial for the clinician and therapist to evaluate how cognitive learning, physical movement and social interaction patterns can be affected with changes in the expressions. For example, in cognitive rehabilitation tasks, subjects tend to make mistakes when there are more negative emotional expressions. Therefore, in such a case, the performance of the subject declines. Similarly, patients are hesitant to involve or sometimes resist to indulge in physiotherapy tasks when they are tired or exhibit negative emotions. In such a scenario, the therapist failed to achieve targets, set for the rehabilitation exercise. In case of social interaction activity it is observed that passive stimulus is required to enhance social interaction, where subjects hardly communicate with other subjects or passively communicate with therapists.

6.1.1 Monitoring negative emotional reactions

Research conducted in [27] illustrates that to achieve the best results, it is essential to determine the emotional states of the patients prior to conducting a rehabilitation exercise. This would have a large impact on an effective rehabilitation as therapists could save time and effort and eventually adapt rehabilitation strategies based on the emotional conditions of the patients. For this purpose, the Pepper robot intervention facilitates the staff members and therapists to determine the emotional states before, during and after the rehabilitation tasks. In addition, Pepper generates reactions according to an individual patient's emotional state to assist in achieving the targets set for the rehabilitation exercise.

6.1.2 Handling negative reactions

In our pilot study, Pepper uses audio, visual and gesture output to handle negative emotional reactions generated by the patients during rehabilitation tasks. The robotic intervention impacted positively on physical rehabilitation but negatively on cognitive activity. In case of physical rehabilitation, patients were motivated to execute more repetitions of tasks. However, patients find the Pepper robot intervention distracting during the cognitive tasks. This is due to the fact that during cognitive activity, Pepper identified their focused-emotional reactions as negative expressions



(a) Static

(b) Dynamic

Fig. 10 ROC curves for emotion recognition through frame-based and sequence of images based information

and reacted accordingly. We implemented a Wizard-of-Oz (WoZ) functionality to recognize behavioral traits in humans to equip the Pepper robot with intellectual cognitive abilities in decision making as well as in creating good relationships with its human user. The WoZ feature aids the therapists to achieve the rehabilitation targets during cognitive task execution and also supports building a reliable relationship between robot and human user.

6.1.3 Performance monitoring

Pepper records each rehabilitation session and generates a pool of expressions over time as illustrated in Fig. 13. The pool of expressions determines the accuracy over the rate of change of expressions from positive to negative and vice versa. In our pilot study, we analyzed subjects exhibit positive expressions while accurately execution of the physical and cognitive tasks. In case of cognitive assessment, pool of expressions are also compared with the results of Android application “Luminosity” that keeps the track of accuracy over the entire session as well as for repetitive tasks for each individual subject. These results also confirmed the exhibition of positive expressions with accuracy of tasks accomplished. During physiotherapy, Pepper acted as a “motivator” that resulted in more repetitions of physical activity during a session for the majority of the patients. The number of robotic reactions in response to positive emotional expressions is directly proportional to the number of repetitions executed in a given session. For instance, in our case study when pepper robot is placed with the subject, number of reps for physiotherapy were increased significantly so the Pepper reactions to acknowledge the effort and motivate the subject. Figure 14 illustrates the Pepper robot interaction with a subject while executing the physiotherapy activity.

6.2 Pepper robot as a feedback agent

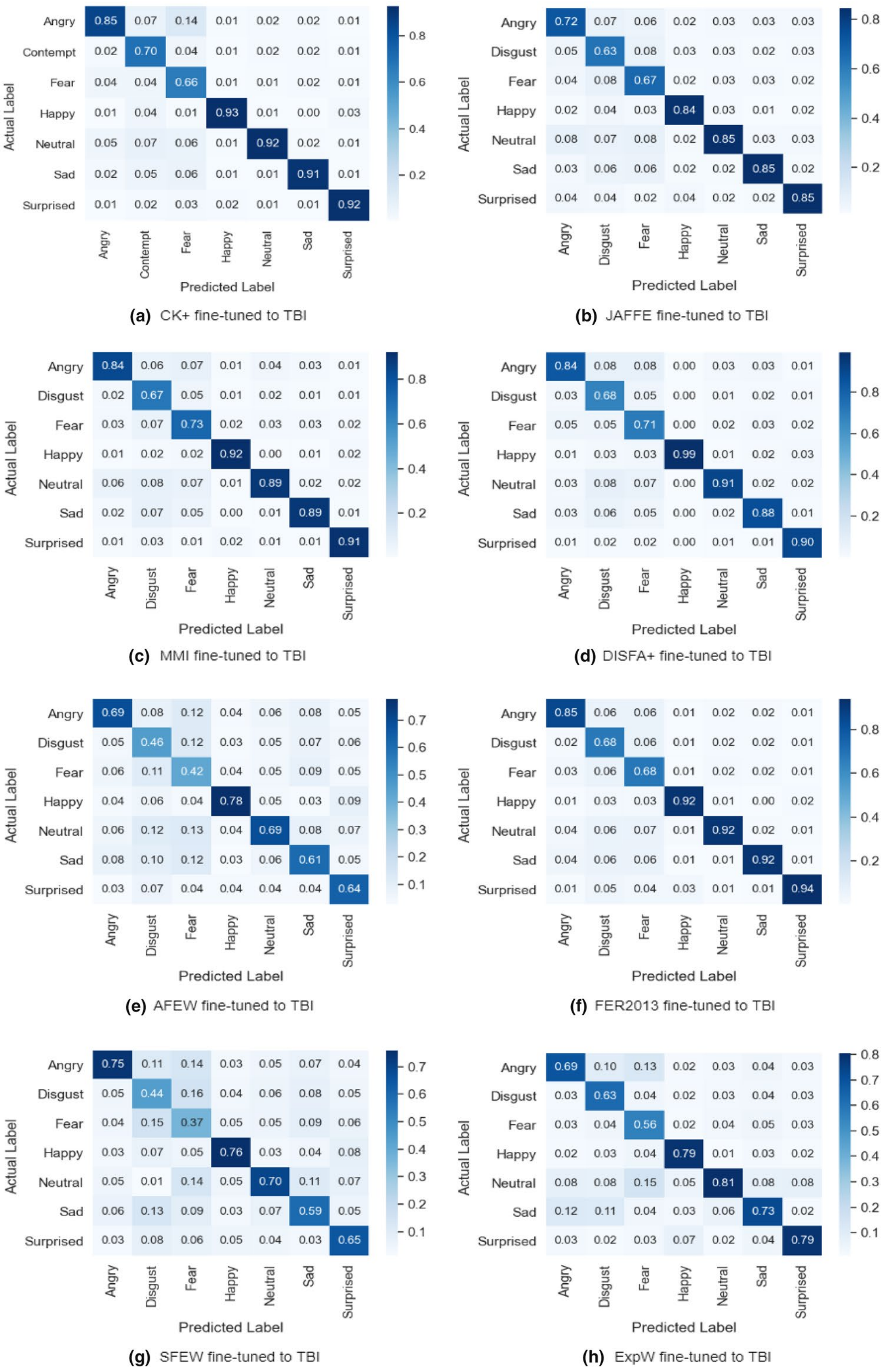
Conventional evaluations involve one-on-one consultations with a therapist. Employing Pepper supports this approach with an objective evaluation of motor and cognitive functions utilizing data obtained during rehabilitation sessions, thus, allowing for accurate, effective, and automated evaluation of motor and cognitive abilities independent of human biases. In addition, audio, visual and gesture output of the Pepper robot during the activity, can provide information about patient-activity-engagement and attention time-span. Attention span of TBI patients is generally low, however, with robotic intervention this issue can be minimized using emotional expression information, where a therapist need to modify the activity to maintain the interest of the subject. This feedback with robotic output and pool of expressions enable the therapists to modify the treatment according to patient involvement and performance.

6.3 Challenges and limitation

We will discuss challenges and limitation related to emotion analysis system and robotic platform and rehabilitation strategies involved as follows.

Comparing the facial expression recognition accuracy with others work is quite challenging as different researchers adopt different databases with varying pre-processing techniques and training techniques. Despite we do performance comparisons with methods explored and average accuracy achieved. We need to consider the balanced and imbalanced data within expression categories for metrics evaluation. Table 7 and Table 8 presents the performance variance of network with varying data classes. Therefore, it is necessary to apply relevant evaluation matrices for system performance analysis.

Although the treatment for rehabilitation through robotic interventions have been proven to be beneficial, in most



(a) CK+ fine-tuned to TBI

(b) JAFFE fine-tuned to TBI

(c) MMI fine-tuned to TBI

(d) DISFA+ fine-tuned to TBI

(e) AFEW fine-tuned to TBI

(f) FER2013 fine-tuned to TBI

(g) SFEW fine-tuned to TBI

(h) ExpW fine-tuned to TBI

Fig. 11 Performance visualization of models fine-tuned to the target TBI database using image frames

facilities they are not yet part of standard care. This is mainly due to the fact that most studies have been carried out with non-mass-developed robotic devices, even though commercially produced social rehabilitation robots are becoming popular, but their costing rise significantly. Along with the need to include more people with clinical rehabilitation substantial attempts are now being made to create and implement low-cost tools that mitigate direct therapist oversight. In the neuro centers, a big obstacle for introducing robot-assisted therapy is that the patient must be able to adhere with the recommended procedure. The patient adherence to recommended treatments in therapy is correlated with

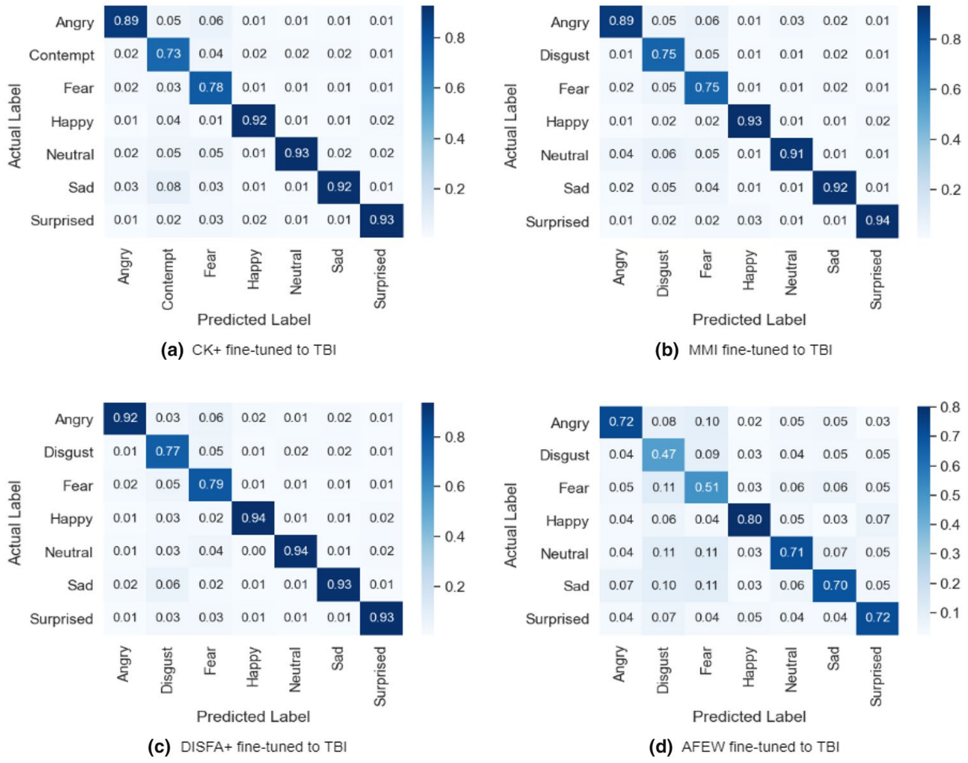


Fig. 12 Performance visualization of models fine-tuned to the target TBI database exploiting temporal information from the image sequences

both decreased compliance and improved treatment outcome. However, lack of desire to do the workouts is one of the key reasons for the inability to adhere. Introduction of more engaging interface such as utilization of the Pepper robot display, synchronized with robotic gestures and audio framework could contribute toward persistent motivation. In addition, where patients impairments are severe, the system can respond by allowing the therapist taking control over the robotic intervention to modify the treatment.

7 Conclusion

In this work we have contributed in two phases, first toward the development of emotion recognition algorithm for TBI patients and second the deployment of the robotic framework for rehabilitation of the TBI patients through the implementation emotion recognition model. For emotion recognition, we have introduced a deep learning framework that is trained to learn the facial features from the datasets acquired in controlled and uncontrolled environment to address two major issues in automatic facial expression recognition. The first problem that we address in this work is non-uniform display of human facial expressions. For instance, in case of TBI patients where facial expressions are variant due to artifacts caused by impairment severity. Employing CNN and CNN-LSTM algorithm, we transferred static and dynamic facial characteristics related to each expressions to TBI patients database having limited identities. One the one hand, our methods have achieved the state-of-the-art performances on specific datasets in both frame-based (static) and sequence of frames-based (dynamic) emotional recognition. Our model has improved the accuracy on various datasets, for instance 78.53% to 79.06% on MMI and 73.73% to 78.19% on FER2013 database in static analysis. Similarly, use of temporal information had enabled the network to exhibit state-of-the-art performance on DISFA and the AFEW with 94.09% and 50.17% accuracy results respectively as presented in Table 6.

On the other hand, our experimental studies reveal that certain facial expressions like anger, fear contempt/disgust, sad and surprise are learnt better from the databases that possess features with frontal faces such as CK+, MMI and JAFFE. Whereas facial features related to neutral, happy expressions have exhibited constant learning pattern in both controlled and in-the-wild environmental conditions. However, large databases in-the-wild like FER2013 and ExpW have produced better results than smaller databases. In addition, posed facial expressions in laboratory or controlled environment, are impure and inconsistent that cause significant degrading in performance of facial expression algorithms in the real world settings. In this work, we train our CNN-LSTM model to transfer facial features in-the-wild settings to the TBI database having pure expressions that were carefully annotated by the experts and clinical staff members, increasing the FER accuracy on the TBI images.

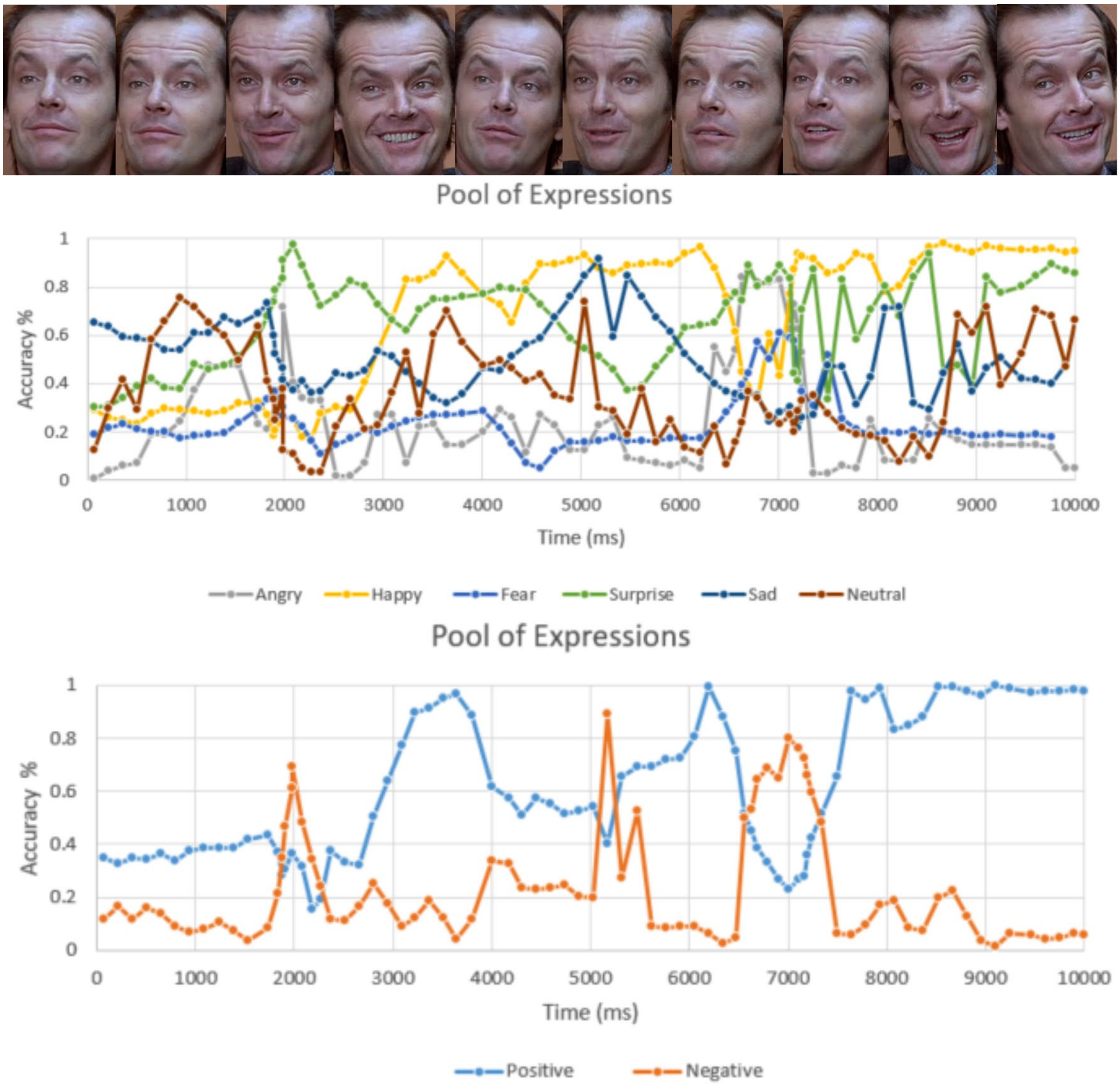


Fig. 13 Visualization of pool of expression in timely order. Video sample of maximum 10 second is taken from AFEW dataset and every 25th frame per second is displayed

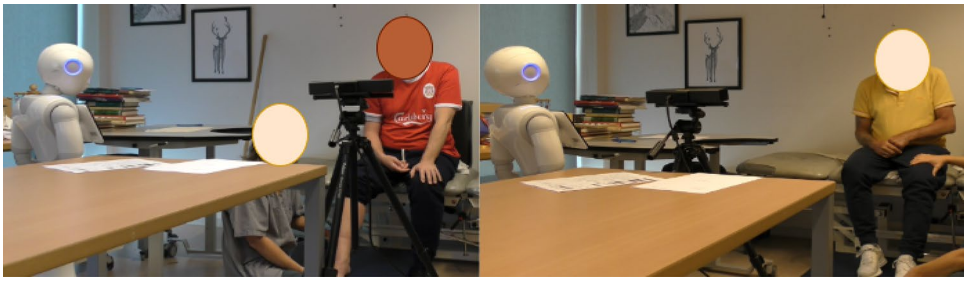


Fig. 14 Visualization of the Pepper robot interaction with the subjects during physical rehabilitation activity. Identities are not covered due to privacy issues

Our experimental findings indicate that the proposed FER algorithm achieves equal or even better performance than state-of-the-art methods.

The second major contribution is the use of robotic technology to transform the recovery from a one-on-one comprehensive care of human beings in specialized institutions to a technologically driven, centrally monitored and controlled environment. Provided the elevated costs associated with long-term recovery and the challenge in maintaining adequate duration and severity of impairment treatment rehabilitation programs, cost-effective deployment of robotic rehabilitation is firmly supported. Implementing emotion understanding through the Pepper robot empowers clinicians to deliver more productive recovery interventions and enable patients to access care more efficiently.

References

1. Adolphs R (2002) Neural systems for recognizing emotion. Curr Opin Neurobiol 12(2):169–177
2. Albiol A, Monzo D, Martin A, Sastre J, Albiol A (2008) Face recognition using hog-ebgm. Pat Recognit Lett 29(10):1537–1543. https://doi.org/10.1016/j.patrec.2008.03.017
3. Albuquerque VHC, Damaševičius R, Garcia NM, Pinheiro PR, et al. (2017) Brain computer interface systems for neurorobotics: methods and applications
4. Balconi M (2012) Neuropsychology of facial expressions. the role of consciousness in processing emotional faces. Neuropsychol Trends 11:19–40
5. Barman A, Chatterjee A, Bhide R (2016) Cognitive impairment and rehabilitation strategies after traumatic brain injury. Indian J Psychol Med 38(3):172–181. https://doi.org/10.4103/0253-7176.183086
6. Bellantonio M, Haque MA, Rodriguez P, Nasrollahi K, Telve T, Escalera S, Gonzalez J, Moeslund TB, Rasti P, Anbarjafari G (2017) Spatio-temporal pain recognition in CNN-based super-resolved facial images. Springer International Publishing, Cham, pp 151–162
7. Bemelmans R, Gelderblom GJ, Jonker P, De Witte L (2012) Socially assistive robots in elderly care: a systematic review into effects and effectiveness. J Am Med Dir Assoc 13(2):114–120
8. Berretti S, Ben Amor B, Daoudi M, del Bimbo A (2011) 3d facial expression recognition using sift descriptors of automatically detected keypoints. Vis Comput 27(11):1021. https://doi.org/10.1007/s00371-011-0611-x
9. Breuer R, Kimmel R (2017) A deep learning perspective on the origin of facial expressions. arXiv preprint arXiv:170501842
10. Burgner-Kahrs J, Rucker DC, Choset H (2015) Continuum robots for medical applications: a survey. IEEE Trans Robot 31(6):1261–1280
11. Cabibihan JJ, Javed H, Ang M, Aljunied SM (2013) Why robots? a survey on the roles and benefits of social robots in the therapy of children with autism. Int J Soc Robot 5(4):593–618
12. Calvaresi D, Cesarini D, Sernani P, Marinoni M, Dragoni AF, Sturm A (2017) Exploring the ambient assisted living domain: a systematic review. J Ambient Intell Human Comput 8(2):239–257
13. Chen L, Zhou M, Su W, Wu M, She J, Hirota K (2018) Softmax regression based deep sparse autoencoder network for facial emotion recognition in human–robot interaction. Inform Sci 428:49–61
14. Corneanu CA, Simón MO, Cohn JF, Guerrero SE (2016) Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications. IEEE Trans Pattern Anal Mach Intell 38(8):1548–1568
15. Dhall A, Goecke R, Lucey S, Gedeon T (2011a) Acted facial expressions in the wild database. Australian National University, Canberra, Australia, Technical Report TR-CS-11 2:1
16. Dhall A, Goecke R, Lucey S, Gedeon T (2011b) Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. In: 2011 IEEE international conference on computer vision workshops (ICCV workshops), IEEE, pp 2106–2112
17. Dhall A, Goecke R, Ghosh S, Joshi J, Hoey J, Gedeon T (2017) From individual to group-level emotion recognition: Emotiw 5.0. In: Proceedings of the 19th ACM international conference on multimodal interaction, pp 524–528
18. Du S, Tao Y, Martinez AM (2014) Compound facial expressions of emotion. Proc Natl Acad Sci 111(15):E1454–E1462. https://doi.org/10.1073/pnas.1322355111
19. Ekman P, Friesen WV, Ellsworth P (2013) Emotion in the human face: guidelines for research and an integration of findings, vol 11. Elsevier, Amsterdam
20. Elaklouk AM, Zin NAM, Shapii A (2015) Investigating therapists' intention to use serious games for acquired brain injury cognitive rehabilitation. J King Saud Univ Comput Inf Sci 27(2):160–169
21. Fan Y, Lu X, Li D, Liu Y (2016) Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In: Proceedings of the 18th ACM international conference on multimodal interaction, pp 445–450
22. Friesen WV, Ekman P, et al. (1983) Emfacs-7: Emotional facial action coding system. Unpublished manuscript, University of California at San Francisco 2(36):1
23. Goodfellow IJ, Erhan D, Carrier PL, Courville A, Mirza M, Hamner B, Cukierski W, Tang Y, Thaler D, Lee DH, et al. (2013) Challenges in representation learning: a report on three machine learning contests. In: International conference on neural information processing, Springer, pp 117–124
24. Guo J, Lei Z, Wan J, Avots E, Hajarolasvadi N, Knyazev B, Kuharenko A, Junior JCSJ, Baró X, Demirel H et al (2018) Dominant and complementary emotion recognition from still images of faces. IEEE Access 6:26391–26403
25. Hassner T, Harel S, Paz E, Enbar R (2015) Effective face frontalization in unconstrained images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4295–4304
26. Hugentobler JA, Vegh M, Janiszewski B, Quatman-Yates C (2015) Physical therapy intervention strategies for patients with prolonged mild traumatic brain injury symptoms: a case series. Int J Sports Phys Therapy 10(5):676
27. Ilyas CMA, Haque MA, Rehm M, Nasrollahi K, Moeslund TB (2018a) Effective facial expression recognition through multimodal imaging for traumatic brain injured patient's rehabilitation. In: International joint conference on computer vision. Springer, Imaging and Computer Graphics, pp 369–389
28. Ilyas CMA, Nasrollahi K, Rehm M, Moeslund TB (2018b) Rehabilitation of traumatic brain injured patients: Patient mood analysis from multimodal video. In: 2018 25th IEEE international conference on image processing (ICIP), IEEE, pp 2291–2295
29. Ilyas CMA, Schmuck V, Haque MA, Nasrollahi K, Rehm M, Moeslund TB (2019) Teaching pepper robot to recognize emotions of traumatic brain injured patients using deep neural networks. In: 28th IEEE international conference on robot and human interactive communication (roman)
30. Jones M, Viola P (2003) Fast multi-view face detection. Mitsubishi Electric Research Lab TR-20003-96 3(14):2
31. Kahou SE, Pal C, Bouthillier X, Froumenty P, Gülçehre Ç, Memisevic R, Vincent P, Courville A, Bengio Y, Ferrari RC, et al. (2013) Combining modality specific deep neural networks for emotion recognition in video. In: Proceedings of the 15th ACM on International conference on multimodal interaction, ACM, pp 543–550
32. Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1725–1732
33. Kaya H, Gürpinar F, Afshar S, Salah AA (2015) Contrasting and combining least squares based learners for emotion recognition

in the wild. In: Proceedings of the 2015 ACM on international conference on multimodal interaction, pp 459–466

34. Kim BK, Dong SY, Roh J, Kim G, Lee SY (2016) Fusing aligned and non-aligned face information for automatic affect recognition in the wild: a deep learning approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 48–57
35. Kim Y, Yoo B, Kwak Y, Choi C, Kim J (2017) Deep generative-contrastive networks for facial expression recognition. arXiv preprint arXiv:170307140
36. King DE (2009) Dlib-ml: a machine learning toolkit. J Mach Learn Res 10(Jul):1755–1758
37. Krishna NM, Sekaran K, Vamsi AVN, Ghantasala GP, Chandana P, Kadry S, Blažauskas T, Damaševičius R (2019) An efficient mixture model approach in brain-machine interface systems for extracting the psychological status of mentally impaired persons using eeg signals. IEEE Access 7:77905–77914
38. Kulkarni K, Corneanu C, Ofodile I, Escalera S, Baro X, Hyniewska S, Allik J, Anbarjafari G (2018) Automatic recognition of facial displays of unfelt emotions. In: IEEE transactions on affective computing
39. Kuo CM, Lai SH, Sarkis M (2018) A compact deep learning model for robust facial expression recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 2121–2129
40. Langhorne P, Bernhardt J, Kwakkel G (2011) Stroke rehabilitation. Lancet 377(9778):1693–1702
41. Li S, Deng W (2020) Deep facial expression recognition: A survey. In: IEEE transactions on affective computing
42. Li S, Deng W, Du J (2017) Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2852–2861
43. Liu M, Wang R, Li S, Shan S, Huang Z, Chen X (2014a) Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild. In: Proceedings of the 16th international conference on multimodal interaction, ACM, pp 494–501
44. Liu P, Han S, Meng Z, Tong Y (2014b) Facial expression recognition via a boosted deep belief network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1805–1812
45. Liu X, Vijaya Kumar B, You J, Jia P (2017) Adaptive deep metric learning for identity-aware facial expression recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 20–29
46. Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: 2010 IEEE computer society conference on computer vision and pattern recognition-workshops, IEEE, pp 94–101
47. Lyons MJ, Akamatsu S, Kamachi M, Gyoba J, Budynek J (1998) The japanese female facial expression (jaffe) database. In: Proceedings of third international conference on automatic face and gesture recognition, pp 14–16
48. Maskeliūnas R, Damaševičius R, Segal S (2019) A review of internet of things technologies for ambient assisted living environments. Future Internet 11(12):259
49. Mavadati M, Sanger P, Mahoor MH (2016) Extended disfa dataset: Investigating posed and spontaneous facial expressions. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 1–8
50. Mavadati SM, Mahoor MH, Bartlett K, Trinh P, Cohn JF (2013) Disfa: a spontaneous facial action intensity database. IEEE Trans Affect Comput 4(2):151–160
51. McKenna K, Cooke DM, Fleming J, Jefferson A, Ogden S (2006) The incidence of visual perceptual impairment in patients with severe traumatic brain injury. Brain Injury 20(5):507–518. https://doi.org/10.1080/02699050600664368
52. Meng Z, Liu P, Cai J, Han S, Tong Y (2017) Identity-aware convolutional neural network for facial expression recognition. In: 2017 12th IEEE international conference on automatic face and gesture recognition (FG 2017), IEEE, pp 558–565
53. Mohammadi MR, Fatemizadeh E, Mahoor MH (2014) Pca-based dictionary building for accurate facial expression recognition via sparse representation. J Vis Commun Image Represent 25(5):1082–1092
54. Müri RM (2016) Cortical control of facial expression. J Comp Neurol 524(8):1578–1585
55. Ng HW, Nguyen VD, Vonikakis V, Winkler S (2015) Deep learning for emotion recognition on small datasets using transfer learning. In: Proceedings of the 2015 ACM on international conference on multimodal interaction, pp 443–449
56. Oquab M, Bottou L, Laptev I, Sivic J (2014) Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1717–1724
57. Otberdout N, Kacem A, Daoudi M, Ballihi L, Berretti S (2019) Automatic analysis of facial expressions based on deep covariance trajectories. In: IEEE transactions on neural networks and learning systems
58. Pantic M, Valstar M, Rademaker R, Maat L (2005) Web-based database for facial expression analysis. In: 2005 IEEE international conference on multimedia and Expo, IEEE, p 5
59. Parkhi OM, Vedaldi A, Zisserman A (2015) Deep face recognition
60. Perry JC, Andureu J, Cavallaro FI, Veneman J, Carmien S, Keller T (2011) Effective game use in neurorehabilitation: user-centered perspectives. In: Handbook of research on improving learning and motivation through educational games: multidisciplinary approaches, IGI Global, pp 683–725
61. Rapple L (2008) Lotsa helping hands. FOCUS J Respirat Care Sleep Med p 36
62. Rees L, Marshall S, Hartridge C, Mackie D, Group MWFTE (2007) Cognitive interventions post acquired brain injury. Brain Injury 21(2):161–200. https://doi.org/10.1080/02699050701201813
63. Robinson H, MacDonald B, Broadbent E (2014) The role of healthcare robots for older people at home: a review. Int J Soc Robot 6(4):575–591
64. Rodil K, Rehm M, Krummheuer AL (2018) Co-designing social robots with cognitively impaired citizens. In: The 10th Nordic conference on human–computer interaction, association for computing machinery
65. Rodriguez P, Cucurull G, Gonzàlez J, Gonfaus JM, Nasrollahi K, Moeslund TB, Roca FX (2017) Deep pain: exploiting long short-term memory networks for facial expression classification. IEEE Trans Cybern 99:1–11
66. Rudovic O, Lee J, Dai M, Schuller B, Picard RW (2018) Personalized machine learning for robot perception of affect and engagement in autism therapy. Sci Robot 3(19):eaao6760
67. Šalkevicius J, Damaševičius R, Maskeliunas R, Laukien I (2019) Anxiety level recognition for virtual reality therapy system using physiological signals. Electronics 8(9):1039
68. Sariyanidi E, Gunes H, Cavallaro A (2014) Automatic analysis of facial affect: a survey of registration, representation, and recognition. IEEE Trans Pat Anal Mach Intell 37(6):1113–1133
69. Shan C, Gong S, McOwan PW (2009) Facial expression recognition based on local binary patterns: a comprehensive study. Image Vis Comput 27(6):803–816. https://doi.org/10.1016/j.imavis.2008.08.005
70. Shapi'i A, Zin M, Azan N, Elaklouk AM (2015) A game system for cognitive rehabilitation. In: BioMed research international 2015

71. Stern Y (2009) Cognitive reserve. Neuropsychologia 47(10):2015–2028
72. Sun N, Li Q, Huan R, Liu J, Han G (2019) Deep spatial-temporal feature fusion for facial expression recognition in static images. Pat Recognit Lett 119:49–61
73. Sutton M (2012) Apps to aid aphasia. ASHA Leader 17(7):32, https://search.proquest.com/docview/1022993653
74. Tang Y (2013) Deep learning using support vector machines. CoRR abs/1306.0239, arXiv:1306.0239
75. Taylor RH, Menciassi A, Fichtinger G, Fiorini P, Dario P (2016) Medical robotics and computer-integrated surgery. In: Springer handbook of robotics, Springer, pp 1657–1684
76. Tian YI, Kanade T, Cohn JF (2001) Recognizing action units for facial expression analysis. IEEE Trans Pat Anal Mach Intell 23(2):97–115
77. Tsaousides T, Gordon WA (2009) Cognitive rehabilitation following traumatic brain injury: assessment to treatment. Mount Sinai J Med J Trans Personal Med 76(2):173–181. https://doi.org/10.1002/msj.20099
78. Uddin MZ, Hassan MM, Almogren A, Alamri A, Alrubaian M, Fortino G (2017) Facial expression recognition utilizing local direction-based robust features and deep belief network. IEEE Access 5:4525–4536
79. Valstar M, Pantic M (2010) Induced disgust, happiness and surprise: an addition to the mmi facial expression database. In: Proceedings of 3rd international workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect, Paris, France, p 65
80. Wan J, Escalera S, Anbarjafari G, Jair Escalante H, Baró X, Guyon I, Madadi M, Allik J, Gorbova J, Lin C, et al. (2017) Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges. In: Proceedings of the IEEE international conference on computer vision, pp 3189–3197
81. Yan J, Zheng W, Cui Z, Tang C, Zhang T, Zong Y (2018) Multi-cue fusion for emotion recognition in the wild. Neurocomputing 309:27–35
82. Yao A, Shao J, Ma N, Chen Y (2015) Capturing au-aware facial features and their latent relations for emotion recognition in the wild. In: Proceedings of the 2015 ACM on international conference on multimodal interaction, ACM, pp 451–458
83. Yao A, Cai D, Hu P, Wang S, Sha L, Chen Y (2016) Holonet: towards robust emotion recognition in the wild. In: Proceedings of the 18th ACM international conference on multimodal interaction, pp 472–478
84. Yin L, Wei X, Sun Y, Wang J, Rosato MJ (2006) A 3d facial expression database for facial behavior research. In: 7th international conference on automatic face and gesture recognition (FGR06), IEEE, pp 211–216
85. Yu Z, Zhang C (2015) Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on international conference on multimodal interaction, pp 435–442
86. Yu Z, Liu Q, Liu G (2018) Deeper cascaded peak-piloted network for weak expression recognition. Vis Comput 34(12):1691–1699
87. Zeng Z, Pantic M, Roisman GI, Huang TS (2008) A survey of affect recognition methods: audio, visual, and spontaneous expressions. IEEE Trans Pat Anal Mach Intell 31(1):39–58
88. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503
89. Zhang K, Huang Y, Du Y, Wang L (2017) Facial expression recognition based on deep evolutional spatial-temporal networks. IEEE Trans Image Process 26(9):4193–4203
90. Zhang X, Yin L, Cohn JF, Canavan S, Reale M, Horowitz A, Liu P (2013) A high-resolution spontaneous 3d dynamic facial expression database. In: 2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG), IEEE, pp 1–6
91. Zhang Z, Luo P, Loy CC, Tang X (2018) From facial expression recognition to interpersonal relation prediction. Int J Comput Vis 126(5):550–569
92. Zhao X, Zhang S (2011) Facial expression recognition based on local binary patterns and kernel discriminant isomap. Sensors 11(10):9573–9588
93. Zhao X, Liang X, Liu L, Li T, Han Y, Vasconcelos N, Yan S (2016) Peak-piloted deep network for facial expression recognition. In: European conference on computer vision, Springer, pp 425–442

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.