Aalborg Universitet

AALBORG UNIVERSITY
DENMARK

Facial Expression Recognition for Traumatic Brain Injured Patients

Ilyas, Chaudhary Muhammad Aqdus; Haque, Mohammad Ahsanul; Rehm, Matthias; Nasrollahi, Kamal; Moeslund, Thomas B.

Published in:
VISAPP 2018

DOI (link to publication from Publisher):
10.5220/0006721305220530

Publication date:
2018

Document Version
Accepted author manuscript, peer reviewed version

Link to publication from Aalborg University

Citation for published version (APA):
Ilyas, C. M. A., Haque, M. A., Rehm, M., Nasrollahi, K., & Moeslund, T. B. (2018). Facial Expression Recognition for Traumatic Brain Injured Patients. In A. Tremeau, J. Braz, & F. Imai (Eds.), *VISAPP 2018: 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications.* (Vol. 4, pp. 522-530). SCITEPRESS Digital Library. https://doi.org/10.5220/0006721305220530

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal -

Take down policy
If you believe that this document breaches copyright please contact us at vbn@aub.aau.dk providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from vbn.aau.dk on: April 24, 2024

Facial Expression Recognition for Traumatic Brain Injured Patients

Chaudhary Muhammad Aqdus Ilyas, Mohammad A. Haque, Matthias Rehm, Kamal Nasrollahi, and Thomas B. Moeslund

Visual Analysis of People Laboratory and Interaction laboratory Aalborg University (AAU), Aalborg, Denmark
{*cmai, mah, matthias, kn, tbm*}*@create.aau.dk*

Keywords: Computer Vision, Face Detection, Facial Landmarks, Facial Expressions, Convolution Neural Networks, Long-Short Term Memory, Traumatic Brain Injured Patients.

Abstract: In this paper, we investigate the issues associated with facial expression recognition of Traumatic Brain Insured (TBI) patients in a realistic scenario. These patients have restricted or limited muscle movements with reduced facial expressions along with non-cooperative behavior, impaired reasoning and inappropriate responses. All these factors make automatic understanding of their expressions more complex. While the existing facial expression recognition systems showed high accuracy by taking data from healthy subjects, their performance is yet to be proved for real TBI patient data by considering the aforementioned challenges. To deal with this, we devised scenarios for data collection from the real TBI patients, collected data which is very challenging to process, devised effective way of data preprocessing so that good quality faces can be extracted from the patients facial video for expression analysis, and finally, employed a state-of-the-art deep learning framework to exploit spatio-temporal information of facial video frames in expression analysis. The experimental results confirms the difficulty in processing real TBI patients data, while showing that better face quality ensures better performance in this case.

1 INTRODUCTION

Facial expression is one of the main sources of communication for human emotions as approximately 55 percent of human communication is happened through facial expressions (Mehrabian, 1968). Computer vision techniques have been developed to extract facial features and use them for different purposes (Mathias et al., 2014) (Klonovs et al., 2016), for example for assessing, mental states (Hyett et al., 2016) (Chen, 2011), health indicators (Li et al., 2012), and various physiological parameters like heartbeat rate, fatigue, blood pressure and respiratory rate (Haque et al., 2016). Among these, automatic detection of facial expression is subject of high importance due to its applications in many fields such as in biometrics, forensics, medical diagnosis, monitoring, defence and surveillance (Ekman and Friesen, 1971) (Mathias et al., 2014) (Du and Martinez, 2015) (Hyett et al., 2016) (Chen, 2011) (Li et al., 2012) (Haque et al., 2015a) (Li and Jain, 2011). Therefore, researchers are putting great emphasis on development of accurate and robust Facial Expression Recognition (FER) systems. A vast body of literature has been produced on this topic in the past decade.

The existing FER systems can be broadly categorized according to their feature extraction methods (Tian et al., 2001) and the used classification techniques. Most widely used methods for facial feature extraction are: geometric features based methods, appearance based methods and hybrid ones (Pantic and Patras, 2006) (Jiang et al., 2014). Geometry based feature extraction methods use geometric shape and position of the facial parts like lips, nose, eyebrows and mouth, with temporal information such as the movement of facial features points from the previous frame to the current frame (Ghimire and Lee, 2013) (Haque et al., 2014). Geometric features are resistant to illumination variation so non-frontal head postures can be handled by processing the figure to frontal head pose to extract the features by measuring distance of fiducial points (Poursaberi et al., 2012) (Anwar Saeed and Elzobi, 2014). Appearance based methods were employed by researchers by using texture information of facial images (Lyons et al., 1999) (li Tian, 2004). In hybrid feature extraction methods both geometric as well as appearance based approaches are deployed for facial image representation (Poursaberi et al., 2012).

FER systems can be further divided on the basis

of classification approaches of extracted facial features. For example, Ghimire et al., 2016 proposed an approach in which both appearance and geometric features are used for facial expression recognition and Support vector Machine (SVM) for classification (Ghimire et al., 2017). Researchers in (Uddin et al., 2017) (Zhao and Zhang, 2011) have used Local Binary Pattern (LBP), Histogram of Oriented Gradient (HoG) in (Ghimire et al., 2017), Linear Discriminant Analysis (LDA) in (Uddin and Hassan, 2015) (Uddin et al., 2017) (Zhao and Zhang, 2011), wavelets based approaches in (Palestra et al., 2015) (Yan et al., 2014) (Poursaberi et al., 2012), Non-Negative Matrix Factorization (NMF) and Discriminant NMF in (de Vries et al., 2015) (Ravichander et al., 2016). Lajevardi and Hussain proposed an investigative analysis on feature extraction and selection models for automatic FER system based on AdaBoost algorithm followed by Gabor filters, log Gabor filters, LBP and higher-order local autocorrelation (HLAC), which is then further modified by applying HLAC-like features (HLACLF) (Lajevardi and Hussain, 2010). Similarly (Ghimire and Lee, 2013) proposed a temporal based FER by tracking the facial feature points and classifying them using multi-class AdaBost and SVM. In (Poursaberi et al., 2012), geometric distance specific fiducial points are determined for FER. Researchers in (Palestra et al., 2015) (Li et al., 2012) (Pantic and Patras, 2006) (Kotsia and Pitas, 2007) used SVM for accurate classification; whereas authors in (Uddin and Hassan, 2015) used the Hidden Markov Models. SVM shows better results when facial expressions are recognized from single frame, but in case of sequence of images HMM produce better results. It is not the set rule as some authors have used combination of different techniques and produced results comparable to state of the art methods.

In recent years more and more researchers have moved towards deep learning techniques for fast, accurate and robust FER. Authors in (Farfade et al., 2015) (Bellantonio et al., 2017) (Triantafyllidou and Tefas, 2016) applied Deep Convolution Neural Network (DCNN) for classification of features into expressions and achieved appreciable results. Yoshihara et al. proposed a feature point detection method for qualitative analysis of facial paralysis using DCNN (Yoshihara et al., 2016). For initial feature point detection, Active Appearance Model (AAM) is used as an input to DCNN for fine tuning. Deep Belief Network (DBN) is another widely used method for robust FER. Kharghanian et al. (Kharghanian et al., 2016) used DBN for pain assessment from facial expressions, where features were extracted with the help of Convolution Deep Belief Networks (CDBN) to identify the pain. Like (Haque et al., 2017), it is tested on the publicly available UNBC McMaster Shoulder Pain database with 95 percentage accuracy. However, these existing methods of FER from healthy people, as used in (Uddin et al., 2017) (Pantic and Patras, 2006) (Kharghanian et al., 2016), are not suitable when applied to real patients in a real scenario.

Recently (Rodriguez et al., 2017) proposed a pain assessment system with FER, where CNN is used to learn facial features from VGG-Faces, then linked to Long Short-Term Memory (LSTM) to take advantage of temporal relations between video frames. This method was further improved by (Bellantonio et al., 2017) by feeding super-resolved facial frames to the CNN+LSTM architecture. These systems of (Rodriguez et al., 2017) and (Bellantonio et al., 2017) work well for extraction of facial expression and its interpretation in form of social signals for healthy people. However, the performance of those systems are yet to be tested on datasets collected on the real patients' scenarios like Traumatic Brain Injured (TBI) patients in a care giving center. This mainly because these patients behavior might be very non-cooperative and non-compliance, and they can have agitation, confusion, loud verbalization, physical aggression, dis-inhibition, impaired reasoning, poor concentration, judgment and mental inflexibility (Lauterbach et al., 2015). Brain injured patients may also have reduced expressions such as smiling, laughing, crying, anger or sadness or their responses may be inappropriate. On the contrary, some TBI patients also exhibit extreme responses like sudden tears, anger outbursts or laughter. It's all due to loss of ability to control over emotions to some extents. These raise the questions whether the state of the art FER systems, like (Rodriguez et al., 2017) and (Bellantonio et al., 2017), will be reliable when working with these patients data. The main issue is that these system require facial images that are good quality and well-posed towards the camera. However, due to the mentioned issues the TBI patients can not always face the camera and their facial images are not of good quality with certainty due to for example rapid changes in head pose. To deal with these difficulties, we equip the state of the art FER system of (Rodriguez et al., 2017) with a Face Quality Assessment (FQA) system that discards most of the faces that are not useful for the FER system and feeds the FER system only with faces that are of better quality compared to the other facial images. We have tested the proposed system on real data of TBI patients which has been collected in a Neurocenter in which these patients are taken care of. To the best of our knowledge, no one has done any previous work on TBI patients to understand their

facial expressions using computer vision techniques. Therefore this work presents a novel experience in this regard and opens up notion for enhancing social communication between patients and care givers.

The rest of this paper is organized as follows. Section 2 describes the proposed methodology for facial feature extraction and recognition of expressions. Section 3 presents the results obtained from the experiments. Finally, Section 4 concludes the paper.

2 THE PROPOSED METHOD

This section describes the architecture of the proposed method for FER analysis in a real patient scenario. The block diagram of the proposed method is illustrated in Figure 1. Following (Rodriguez et al., 2017), in the first step, the face is detected from a input video. In order to reduce erroneous detection of face we employ a face alignment approach by detecting facial landmarks. The detected landmarks are tracked and faces are cropped according to the landmark positions. In the next step face quality is assessed by following (Haque et al., 2015b) and only good quality faces are stored in face log. Faces are then fed to a CNN. This network was pre-trained with VGG-16 faces as used by (Bellantonio et al., 2017) and (Rodriguez et al., 2017). These steps of the system are further explained in the following subsections.

2.1 Data Acquisition and Preprocessing

The subjects are filmed by a Axis RGB-Q16 camera with resolution of 1280 x 960 to 160 x 90 pixels at 30fps (frames per second). Then, these images are fed to a facial image acquisition system which consists of three steps: face detection, face quality assessment and face logging. The first step is face detection from the video frames for which we used a well-know method, called VJ (Viola and Jones) face detector (Viola and Jones, 2001). Due to its speed and moderately high accuracy by using Haar-like features we selected this method. This method constructs a classifier with the help of learning algorithm based on AdaBoost which effectively classify the images on the basis of few critical features from large set and discard background regions by cascading. However, it is prone to erroneous detection when face is in low quality in terms of occlusion or pose variation. On the other hand, while most FER databases have near-frontal head poses of good quality images with very less occlusions (no spectacles, hand gestures covering the mouth, etc.), in our case subjects are TBI patients and they are not cooperative enough to ensure good facial data capturing. So there is high possibility of non-frontal view and continuous pose variations, resulting in low quality of images and consequently large amount of miss detected faces as shown in Figure 2. Moreover, due to inability to recognize and appropriately respond to non-verbal cues TBI patients have feeble response (Bird and Parente, 2014). This in turns increases the complexity of data collection. Thus, instead of detecting face in every single frames of a video, we employ a face alignment method on a properly detected face frame in the video and then track the facial landmarks in the subsequent frames. This reduces the possibility of erroneous detection by VJ in subsequent video frames, as the face is tracked instead of detected again and again in the video sequence.

Face alignment is a process of localization of inner facial structures such as apex of the nose or curve of the eye by using some predefined landmarks that help in better enrolment of the face. Such land-marking also helps in the speedy extraction of geometric structures as well as additional strong local characteristics. Due to advancement in technology, regression based facial land-marking methods have contributed towards the automatic face alignment. One of the most effective approach is the Supervised Decent Method (SDM) (Xiong and la Torre, 2013). In SDM, 49 facial landmarks are applied around eyes corners, nose line, lips and eye borrows. In addition, SDM uses small optical flow vectors and pixel by pixel neighbourhood measure by avoiding window based point tracing. This provides high computational efficiency, and more stable and precise tracking for long time period of visual frames as demonstrated by (Haque et al., 2015b). Thus, we employ the SDM based face alignment in the proposed method of FER. The steps of face alignment in a video is shown in Figure 3. The face is first detected in a video frame by an off-the-shelf face detector (VJ in this case) and then the facial landmarks are identified in that frame. Instead of detecting face in the subsequent video frames, those landmarks are tracked in the subsequent frames. The performance of following the SDM-based approach over mere VJ will be evaluated in the experimental result section. By using the landmarks, we find the face boundary and then crop the faces. The faces are then forwarded to the next step.

2.2 Face Quality Assessment

System performance for FER is highly dependent on the quality of facial images. In practice for the TBI

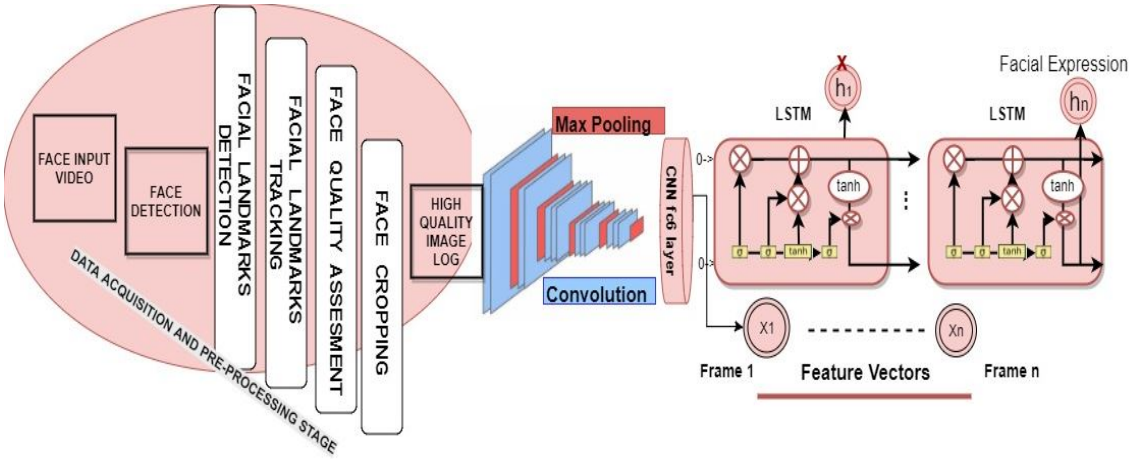

Figure 1: Block diagram of Facial Expression Recognition System based on CNN+LSTM model to exploit spatio-temporal information

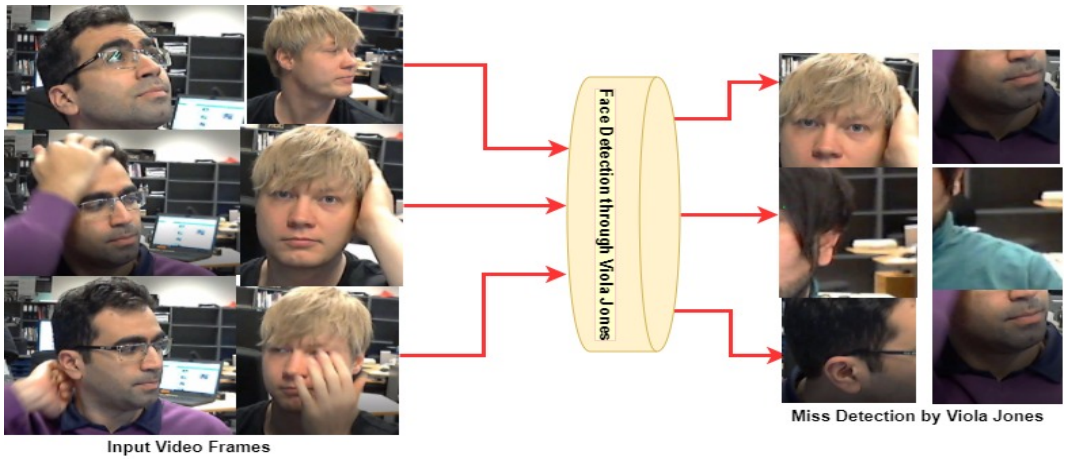

Figure 2: Miss detection of faces by VJ face detector due to occlusion or high pose variation

patient dataset, there is high possibility of non-frontal view of face and continuous pose variations, resulting in low quality of images, even though those faces are tracked by the SDM. Figure 4 show the case of occluded face (which of course means low quality) for a video sequence where average pixel intensities are varying due to the presence and absence of occlusion over time. To avoid such problems, we employ a FQA technique on the faces cropped after SDM. This is accomplished by measuring some face quality matrices like image resolution, sharpness, and face rotation as shown in (Haque et al., 2013). Before logging facial frames into final face log for FER, low quality face frames are identified by setting first frame as a reference frame and comparing similarity in the rest of the frames in a particular event of video as follow in (Irani et al., 2016). Similarity of frames is calculated by the following equation:

$$S_{Clr} = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N}(\mathbf{A}_{mn}-\overline{A})(\mathbf{B}_{mn}-\overline{B})}{\sqrt{\sum_{m=1}^{M}\sum_{n=1}^{N}(\mathbf{A}_{mn}-\overline{A})^2\sum_{m}^{M}\sum_{n}^{N}(\mathbf{B}_{mn}-\overline{B})^2}} \tag{1}$$

In the above equation A and B are the reference faces whereas $\overline{\mathbb{A}}$ and $\overline{\mathbb{B}}$ are average pixels levels of the current frame. M and N are number of rows and columns in an image matrix. The degree of dissimilarity calculated from the above equation forms the basis for face quality score. The more the dissimilarity the more the possibility of a low quality face.

2.3 Face logging

In this step, the faces obtained after SDM tracking are considered along with their associated quality score. If the score is lower than a predefined threshold we

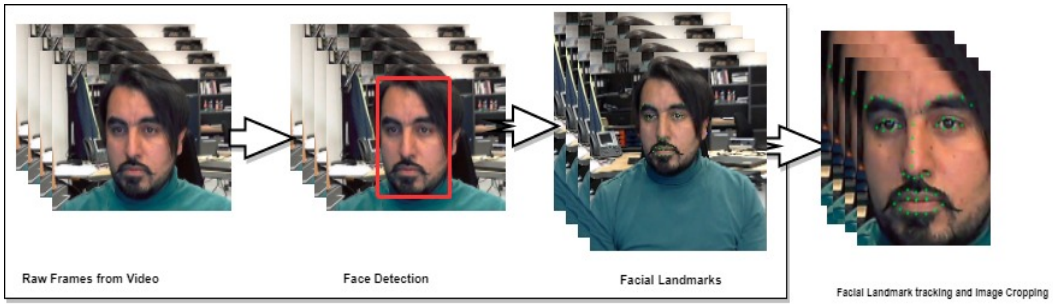

Figure 3: Facial landmark identification and tracking in Supervised Descent Method (SDM)

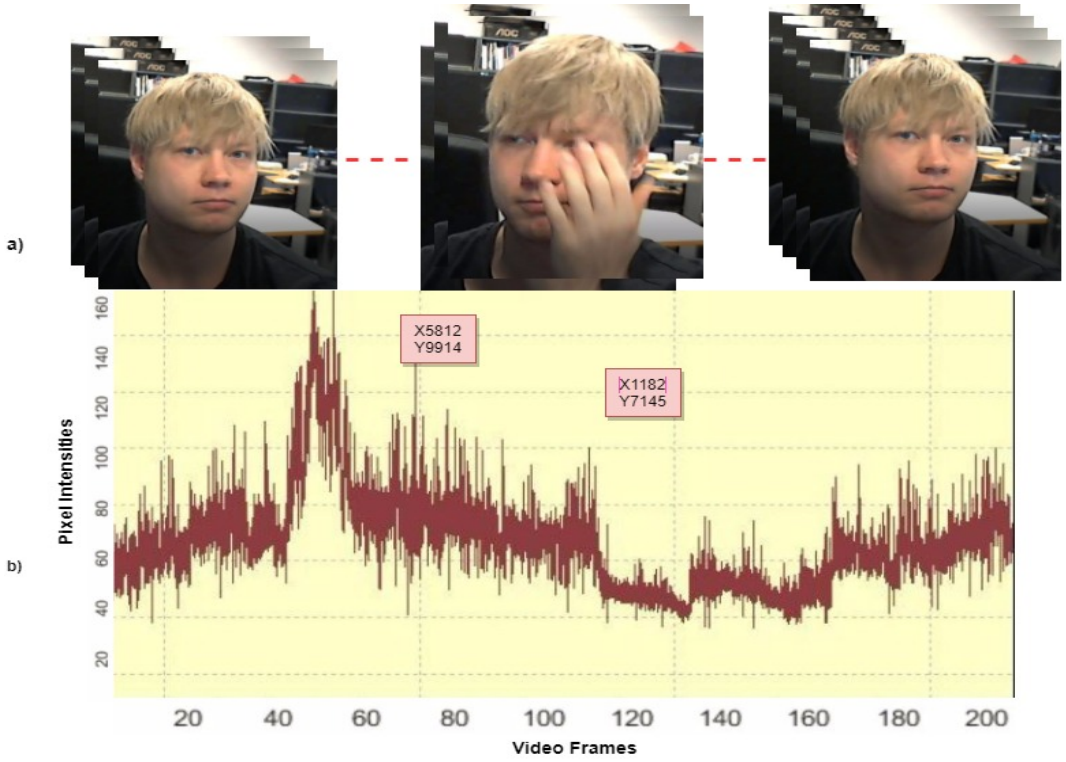

Figure 4: Depiction of varying pixels intensities due to the presence and absence of occlusion over time. a) shows the example face frames and b) show the variation in pixel intensities over time

simply discard that before logging. Once the quality of face is ensured, images are cropped to a common input size of neural network (224x224 pixels in our experiment) and these are ready to feed to the deep learning architecture.

2.4 The CNN+LSTM based Deep Learning Architecture for FER

Convolutional neural networks are specialized set of neuron networks having multiple layers of input and output that utilizes the local features in image to obtain the visual information. CNN has multiple layers for convolution and padding. A typical 2-Dimensional (2D) CNN takes 2D images as input and considers each image as a *nn* matrix. Generally, parameters of the CNN are randomly initialized and learned by performing gradient descend using a back propagation algorithm. It uses a convolution operator in order to implement a filter vector. The output of the first convolution will be a new image, which will be passed through another convolution by a new filter. This procedure will continue until the most suitable feature vector elements $\{V_1, V_2, ..., V_n\}$ are found. Convolutional layers are normally alternated with another type of layer, called Pooling layer, which function is to reduce the size of the input in order to reduce the spatial dimensions and gaining computational performances and translation invariance (Noroozi et al., 2017). CNN performed remarkably well in facial recognition (Ji et al., 2013) as well as automatic fa-

cial detection (Farfade et al., 2015). In order to take advantage of its good results for FER we have applied this method on TBI patients data to extract facial features relevant to FER.

In general, CNN deals with images that are isolated. However, in our case we have used the sequences of images in a timely manner and thus, having the notion of using temporal information as well. So to exploit the temporal information associated with facial expression in video, we have used an implementation of Recurrent Neural Network (RNN), that is capable of absorbing the sequential information, called LSTM model from (Rodriguez et al., 2017). The LSTM states are controlled by three gates associated with forget (f), input (i), and output (o) states. These gates control the flow of information through the model by using point-wise multiplications and sigmoid functions σ, which bound the information flow between zero and one by the followings:

$$i(t) = \sigma(W_{(x\to i)}x(t) + W_{(h\to i)}h(t-1) + b_{(1\to i)}) \quad (1)$$

$$f(t) = \sigma(W_{(x\to f)}x(t) + W(h\to f)h(t-1) + b_{(1\to f)}) \quad (2)$$

$$z(t) = tanh(W_{(x\to c)}x(t)) + W_{(h\to c)}h(t-1) + b_{(1\to c)}) \quad (3)$$

$$c(t) = f(t)c(t-1) + i(t)z(t), \quad (4)$$

$$o(t) = \sigma(W_{(x\to o)}x(t) + W_{(h\to o)}h(t-1) + b_{(1\to o)}) \quad (5)$$

$$h(t) = o(t)tanh(c(t)), \quad (6)$$

where $z(t)$ is the input to the cell at time t, c is the cell, and h is the output. $W_{(x\to y)}$ are the weights from x to y.

In this paper, we use a combination of CNN and LSTM where CNN extract facial features from the faces logged from the TBI patients video and LSTM find temporal correlation based on those features in temporal setting. A schematic diagram of the CNN+LSTM is shown in the right hand side of the Figure 1 and more details can be found in (Rodriguez et al., 2017). We used a off-the-shelf fine-tuned version of the VGG-16 CNN model (Parkhi et al., 2015) pre-trained with faces for spatial feature extraction. We obtained the features of the `fc7` layer of the CNN (VGG-16) and then use them as input to a the LSTM to exhibit hybrid deep learning performance by CNN+LSTM. The implementation of the CNN+LSTM is available online through (Rodriguez et al., 2017).

3 EXPERIMENTAL RESULTS

In this section, we first describe the database captured and used during our investigation. We then demonstrate and commented on the results.

3.1 The Database

In order to have experiments for FER on TBI patients data, we require a database. However, to the best of our knowledge, there is no publicly available facial video database from real TBI patients. In establishment of a database, first task was identification of data collection methods. Most of TBI patients have varying ability to identify and respond to non-verbal expression of emotions (Bird and Parente, 2014). After visiting different neurocenters and care-homes where TBI patients are provided rehabilitation facilities around Denmark, and consulting with experts and care-givers who are in direct contact with TBI patients, we have finalized three uniform scenarios for data collection from all the patients under observation. The uniformity in data collection is maintained to have reliable data for future use. Those scenarios are: a) cognitive rehabilitation therapy, b) physiotherapy, and c) social communication with other residents of the neurocenter. In cognitive therapy, a TBI patient plays a game or mind quiz in order to judge how much thinking or cognitive ability a particular subject posses. On the basis of this activity further data elicitation process is organized. In the second activity of physiotherapy, subjects stress level of fatigue is determined. The last activity, where TBI patients have to interact with other patients and care-givers, provides insight about patient ability to give and perceive communication signals.

On contrary to normal people, TBI patients have intolerance, rapid mood swings accompanied by anger or tear bursts, low concentration and impaired facial emotion recognition. Considering these challenges, collection of data, particularly facial videos, is not a trivial task as most of patients do not keep their face positions still. Even if they do so, it is still not easy to understand their emotions for some other problems. Mostly they have sad or depressed emotions after post traumatic life. However, experts who are dealing with TBI patients over certain period of time are able to annotate the patients emotional status as neutral or normal expression. Another problem is: they get agitated very quickly and so it was big task to involve them in the aforementioned three activities. For this purpose, to have clear and precise emotion recognition, we devised a game in such a way that we intentionally let the patients to win to see their happy expressions. Similarly to have their head posed in front of camera, a tablet displaying emotional scenes, is placed just parallel to camera recording their facial expressions. Similar adjustments are

Table 1: The performance of SDM-based face alignment and tracking to extract faces from the video frames in comparison to basic VJ face detector.

Number of Frames	**VJ**	**SDM**
Total no. of frames	27689	25289
Training frames	22082	20403
Testing frames	5607	4886
Total mis-detection	2429	1128
Percentage Error	**8.67 %**	**4.46 %**

made in other activities during recording. One interesting observation is that all the TBI patients have taken deep interest in mind game, and movie or picture illustration regardless of their disability nature. This allows us to collect more neutral, happy and angry expressions. However, we could not collect much expressions of sadness, surprised and fatigue due to non-cooperation, traumatic disabilities and other social and technical issues.

We collected data in multiple phases in a number of sessions. In total we got 539 video sequences (one sequence means one expression event) with variable lengths (1-5 seconds). However, we observe that the data is highly imbalanced as out of 539 events 463 are of neutral expression. In other words, out of approximately 20,000 frames, almost 14000 represents neutral expressions. Among others, 108 events (app. 3300 frames) of happy, 72 events (app. 2200 frames) of angry and very few are other expressions. On other hand, most of them have too much head motions, so making the data even more challenging for further processing.

3.2 Performance Evaluation

In this section, we first demonstrate the impact of employing a SDM-based face aligner and tracker over VJ face detector. Table 1 shows the amount of erroneous face detection in the video frames. From the results, we observe that FQA removed 2429 erroneous faces out of 27689 while using VJ. It means that 8.67 percentage of the detection were not correct by VJ. On other hand, when FQA technique is employed on faces detected by SDM, 4.46 percentage of the facial frames were not detected correctly as FQA discarded 1128 frames out of 25289 frames. Comparing both results, SDM-based detection by using alignment and tracking provided better accuracy in finding the right faces.

Table 2 shows the accuracy of FER in terms of AUC for two scenarios while the number of epochs in the LSTM was varying in yielding the results. The epochs of CNN-LSTM system is gradually increased

Table 2: AUC results for FER of TBI patients data with gradual increase in epoch values.

Area Under Curve (AUC)		
Epocs Value	**Viola Jones**	**SDM**
10	66.37	69.49
15	69.55	72.03
20	75.42	63.21
25	76.94	72.96
30	67.31	75.26
35	75.76	72.35
40	63.03	73.38
45	67.08	71.81
50	68.63	74.85

Table 3: AUC results for FER of TBI patients data with gradual increase in RHO values.

	Area Under Curve (AUC)		
RHO Values	**Full Frames**	**Viola Jones**	**SDM**
1	51.43	61.18	70.17
3	54.26	62.08	72.09
5	53.21	63.03	73.38
7	59.27	63.57	72.27
9	57.12	64.5	72.83
11	59.17	63.29	71.09

by step of 5, from 5 to 50 keeping other parameters such as RHO, recurrent depth, and drop-out probability constant. From the results we observe that the accuracy of VJ-based CNN-LSTM system is increased with gradual increase in epochs up to 25 epochs. It reached up to level of 76.94 percent at the 25th epoch. At 30th epoch, its value was dropped down to 67.31 percent, but strangely jumped to 75.26 percent in a higher values of epochs.

Table 3 show the effect of chaning RHO value for three scenarios. From the results we observe that the SDM-based approach reached maximum AUC value of 75.26 percent. RHO value is gradually changed at step of 2, from 1 to 11, means giving more temporal information for FER, while keeping the epochs constant. AUC values showed the VJ-based approach exhibits slightly higher accuracy by increasing temporal information. In contrast, SDM-based approach got the accuracy above 70 percent in all steps with maximum value of 73.38 percent and minimum value of 70.17 percent. Similar uphill trends is observed up to RHO 5 and then a slight decline is observed.

It is clearly evident from the experiment results for TBI patients data, despite of the challenging datasets accuracy of system is increased to certain extent.

4 CONCLUSIONS

In this paper, we pointed out the rationale about investigating facial expression analyzing system by using data obtained from real TBI patients. The study reveals the challenges associated with real-world scenarios including patients, instead of healthy volunteers used in the previous works. We captured data from TBI patients in a neurocenter, extracted faces from the video frames by employing different methods to find out the effective one. We then fed the cropped faces into a CNN+LSTM based deep learning framework to exploit both spatio-temporal information to detect the patients mental status in terms of facial expressions. The results were demonstrated with different spatio-temporal parameters of the system. The result showed that the facial information obtained from patient is varying in such a way that it is hard to predict the expression with high accuracy. Moreover, we observed strong effect of employing an effective face detection method with face quality assessment for FER. However, as a note for future work, further processing such as face frontalization, larger dataset for training and subject specific knowledge base incorporation might be useful in improving the performance.

REFERENCES

Anwar Saeed, Ayoub Al-Hamadi, R. N. and Elzobi, M. (2014). Frame-based facial expression recognition using geometrical features. *Advances in Human-Computer Interaction*, 2014:1–13.

Bellantonio, M., Haque, M. A., Rodriguez, P., Nasrollahi, K., Telve, T., Escalera, S., Gonzalez, J., Moeslund, T. B., Rasti, P., and Anbarjafari, G. (2017). *Spatio-temporal Pain Recognition in CNN-Based Super-Resolved Facial Images*, pages 151–162. Springer International Publishing, Cham.

Bird, J. and Parente, R. (2014). *Recognition of nonverbal communication of emotion after traumatic brain injury*.

Chen, Y. (2011). *Face Perception in Schizophrenia Spectrum Disorders: Interface Between Cognitive and Social Cognitive Functioning*, pages 111–120. Springer Netherlands, Dordrecht.

de Vries, G.-J., Pauws, S., and Biehl, M. (2015). *Facial Expression Recognition Using Learning Vector Quantization*, pages 760–771. Springer International Publishing, Cham.

Du, S. and Martinez, A. M. (2015). Compound facial expressions of emotion: from basic research to clinical applications. *Dialogues Clin Neurosci*, 17(4):443–455.

Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *J Pers Soc Psychol*, 17(2):124–129.

Farfade, S. S., Saberian, M. J., and Li, L.-J. (2015). Multi-view face detection using deep convolutional neural networks. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, ICMR '15, pages 643–650. ACM.

Ghimire, D. and Lee, J. (2013). Geometric feature-based facial expression recognition in image sequences using multi-class adaboost and support vector machines. *Sensors*, 13(6):7714–7734.

Ghimire, D., Lee, J., Li, Z.-N., and Jeong, S. (2017). Recognition of facial expressions based on salient geometric features and support vector machines. *Multimedia Tools and Applications*, 76(6):7921–7946.

Haque, M. A., Irani, R., Nasrollahi, K., and Moeslund, T. B. (2016). Facial video-based detection of physical fatigue for maximal muscle activity. *IET Computer Vision*, 10(4):323–329.

Haque, M. A., Nasrollahi, K., and Moeslund, T. B. (2013). Real-time acquisition of high quality face sequences from an active pan-tilt-zoom camera. In *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 443–448.

Haque, M. A., Nasrollahi, K., and Moeslund, T. B. (2014). Constructing facial expression log from video sequences using face quality assessment. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 2, pages 517–525.

Haque, M. A., Nasrollahi, K., and Moeslund, T. B. (2015a). *Heartbeat Signal from Facial Video for Biometric Recognition*, pages 165–174. Springer International Publishing, Cham.

Haque, M. A., Nasrollahi, K., and Moeslund, T. B. (2015b). Quality-aware estimation of facial landmarks in video sequences. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 678–685.

Haque, M. A., Nasrollahi, K., and Moeslund, T. B. (2017). Pain expression as a biometric: Why patients' self-reported pain doesn't match with the objectively measured pain? In *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, pages 1–8.

Hyett, M. P., Parker, G. B., and Dhall, A. (2016). *The Utility of Facial Analysis Algorithms in Detecting Melancholia*, pages 359–375. Springer International Publishing, Cham.

Irani, R., Nasrollahi, K., Dhall, A., Moeslund, T. B., and Gedeon, T. (2016). Thermal super-pixels for bimodal stress recognition. In *2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6.

Ji, S., Xu, W., Yang, M., and Yu, K. (2013). 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231.

Jiang, B., Valstar, M., Martinez, B., and Pantic, M. (2014). A dynamic appearance descriptor approach to facial

actions temporal modeling. *IEEE Transactions on Cybernetics*, 44(2):161–174.

Kharghanian, R., Peiravi, A., and Moradi, F. (2016). Pain detection from facial images using unsupervised feature learning approach. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 419–422.

Klonovs, J., Haque, M. A., Krueger, V., Nasrollahi, K., Andersen-Ranberg, K., Moeslund, T. B., and Spaich, E. G. (2016). *Monitoring Technology*, pages 49–84. Springer International Publishing, Cham.

Kotsia, I. and Pitas, I. (2007). Facial expression recognition in image sequences using geometric deformation features and support vector machines. *IEEE Transactions on Image Processing*, 16(1):172–187.

Lajevardi, S. and Hussain, Z. (2010). Novel higher-order local autocorrelation-like feature extraction methodology for facial expression recognition. *IET Image Processing*, 4:114–119(5).

Lauterbach, M. D., Notarangelo, P. L., Nichols, S. J., Lane, K. S., and Koliatsos, V. E. (2015). Diagnostic and treatment challenges in traumatic brain injury patients with severe neuropsychiatric symptoms: insights into psychiatric practice. *Neuropsychiatr Dis Treat*, 11:1601–1607.

Li, F., Zhao, C., Xia, Z., Wang, Y., Zhou, X., and Li, G.-Z. (2012). Computer-assisted lip diagnosis on traditional chinese medicine using multi-class support vector machines. *BMC Complementary and Alternative Medicine*, 12(1):127.

Li, S. Z. and Jain, A. K. (2011). *Handbook of Face Recognition*. Springer Publishing Company, Incorporated, 2nd edition.

li Tian, Y. (2004). Evaluation of face resolution for expression analysis. In *2004 Conference on Computer Vision and Pattern Recognition Workshop*, pages 82–82.

Lyons, M. J., Budynek, J., and Akamatsu, S. (1999). Automatic classification of single facial images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(12):1357–1362.

Mathias, M., Benenson, R., Pedersoli, M., and Van Gool, L. (2014). *Face Detection without Bells and Whistles*, pages 720–735. Springer International Publishing.

Mehrabian, A. (1968). Communication without words. *Psychology Today*, 1.2(4):53–56.

Noroozi, F., Marjanovic, M., Njegus, A., Escalera, S., and Anbarjafari, G. (2017). Audio-visual emotion recognition in video clips. *IEEE Transactions on Affective Computing*, (99):1–1.

Palestra, G., Pettinicchio, A., Del Coco, M., Carcagnì, P., Leo, M., and Distante, C. (2015). *Improved Performance in Facial Expression Recognition Using 32 Geometric Features*, pages 518–528. Springer International Publishing, Cham.

Pantic, M. and Patras, I. (2006). Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 36(2):433–449.

Parkhi, O. M., Vedaldi, A., and Zisserman, A. (2015). Deep face recognition. In *British Machine Vision Conference*.

Poursaberi, A., Noubari, H. A., Gavrilova, M., and Yanushkevich, S. N. (2012). Gauss–laguerre wavelet textural feature fusion with geometrical information for facial expression identification. *EURASIP Journal on Image and Video Processing*, 2012(1):17.

Ravichander, A., Vijay, S., Ramaseshan, V., and Natarajan, S. (2016). *Automated Human Facial Expression Recognition Using Extreme Learning Machines*, pages 209–222. Springer International Publishing, Cham.

Rodriguez, P., Cucurull, G., Gonzlez, J., Gonfaus, J. M., Nasrollahi, K., Moeslund, T. B., and Roca, F. X. (2017). Deep pain: Exploiting long short-term memory networks for facial expression classification. *IEEE Transactions on Cybernetics*, PP(99):1–11.

Tian, Y. I., Kanade, T., and Cohn, J. F. (2001). Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115.

Triantafyllidou, D. and Tefas, A. (2016). Face detection based on deep convolutional neural networks exploiting incremental facial part learning. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3560–3565.

Uddin, M. Z. and Hassan, M. M. (2015). A depth video-based facial expression recognition system using radon transform, generalized discriminant analysis, and hidden markov model. *Multimedia Tools and Applications*, 74(11):3675–3690.

Uddin, M. Z., Hassan, M. M., Almogren, A., Alamri, A., Alrubaian, M., and Fortino, G. (2017). Facial expression recognition utilizing local direction-based robust features and deep belief network. *IEEE Access*, 5:4525–4536.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–511–I–518 vol.1.

Xiong, X. and la Torre, F. D. (2013). Supervised descent method and its applications to face alignment. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 532–539.

Yan, J., Zhang, X., Lei, Z., and Li, S. Z. (2014). Face detection by structural models. *Image and Vision Computing*, 32(10):790 – 799. Best of Automatic Face and Gesture Recognition 2013.

Yoshihara, H., Seo, M., Ngo, T. H., Matsushiro, N., and Chen, Y. W. (2016). Automatic feature point detection using deep convolutional networks for quantitative evaluation of facial paralysis. In *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 811–814.

Zhao, X. and Zhang, S. (2011). Facial expression recognition based on local binary patterns and kernel discriminant isomap. *Sensors*, 11(10):9573–9588.