ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/340133756

Modified CNN-LSTM for Pain Facial Expressions Recognition

Article · March 2020

CITATIONS
0

READS
125

3 authors, including:

Ibraheem Nadher Ibraheem
Al-Mustansiriya University
10 PUBLICATIONS **4** CITATIONS

SEE PROFILE

All content following this page was uploaded by Ibraheem Nadher Ibraheem on 25 March 2020.

The user has requested enhancement of the downloaded file.

International Journal of Advanced Science and Technology
Vol. 29, No. 3s, (2020), pp. 304-312

Modified CNN-LSTM for Pain Facial Expressions Recognition

Wafaa M. Salih Abedi[1], Dr. Ahmed T. Sadiq[1], Dr. Ibraheem Nadher[2]

[1]*Computer Science Department, University of Technology, Baghdad, Iraq*
[2]*Faculty of Basic Education, AL- Mustansiriya University, Baghdad, Iraq*

Abstract

Computer vision studies provide efficient methods for facial expressions recognition. As soon as deep machine learning approaches became involved in producing an automatic facial expressions recognition revealed even enhanced performance. In spite of the improvements have already existed in this field, researches are still need enhancements toward pain expressions recognition. To deal with the gap in existing researches, the proposed method is a hybrid model consists of a combination of the Convolutional Neural Network (CNN) and Long-Short Term Memory (LSTM) model. The proposed method contribution is to successfully detect and classify the pain expression from a video sequence. The proposed hybrid model utilizes three significant factors: the spatial data presented for each frame, temporal data for each pain expression pattern, and variant face resolution. The proposed hybrid combination examined on the widely present UNBC-McMaster Shoulder Pain dataset. The results of the proposed process success to detect the pain expressions for the video sequence with an accuracy of 73.31%.

CNN+LSTM模型使用了三个主要因素：
1）单帧的控件数据
2）痛苦表情模式的时域数据
3）面部分辨率的多样性

测试数据集：
UNBC-McMaster Shoulder Pain dataset

Keywords: *Long short term memory (LSTM); image classification; face expression recognition; convolutional neural network (CNN); Deep learning;*

1. Introduction

The notion of computer vision-based automatic pain level assessment introduced, since the traditional methods for pain recognition measured by self-monitoring [1], which needs specific knowledge, and abilities, and it should be managed efficiently as professional and moral responsibilities [2]. These entire aspects make self-monitoring impossible for older patients and children [3]. The challenges increase in the existence of external features like gender difference [4].

Pain expressions recognition is a part from face expressions recognition. Recently, advance images/videos analysis by the use of deep learning structures such as recurrent neural networks (RNN) or convolutional neural networks (CNN) architectures for facial expressions recognitions [5] and emotions detections [6].

Deep learning, while applying its methods to learn the data that is useful to images classifications, suggests some progressive classifications technology that supports the control of covering many layers [7].

The algorithms are trained first with lots of known databases and after that; the network is experienced using the remaining portions of the datasets. This process delivers the feature maps that effect the correct image classifications [8].

The limitation of network arrangements and the lack of suitable datasets for training, the network is unable to produce correctly the required feature maps, and this might be the result of networks over-fitting. To soften this problem, proper RNNs classifiers with inventive deep learning algorithms are used [9].

A deep learning framework for pain assessments presented in [10], involves two types of data from videos sequences: spatial data provide facial expression data in a single frame, and temporal data shows the connection between the expressions in successive frames. The image resolution from the imaging acquisition device or sensor is the first

ISSN: 2005-4238 IJAST
Copyright ⓒ 2020 SERSC

limitation. The solution was to increase the sensors density and reduce the sensors size, which may cause the snapshot noise [11].

This paper study the reasonableness of the use of CNN to extract the spatial features, and fed these features to Long Short-Term Memory (LSTM), which achieve the temporal data from each video frame to estimate the pain expressions. The result obtained through the use of widely accessible challenging dataset named UNBC-McMaster Shoulder Pain dataset [12].

The remaining sections of this paper organized as follows: Section 2 presents an indication of related works; Section 3 presents the RNNs concept. Section 4 presents the proposed method. Section 5 discusses the experimental dataset and result. Section 6 summarizes and concludes.

2. Related Work

In computer vision, pain recognition is an important process in the direction of an automatic finder of facial expressions. The UNBC-McMaster dataset frames used to suggest new methods for pain recognition. Lucey et al. [13], proposed method of using support vector machine (SVM) and active appearance models (AAMs) as a baseline with UNBC-McMaster dataset [14] to calculate pain actions. Rudovic et al. [15], estimate the intensity of face expressions depending on the context, such as head movement and illuminations.

Valstar and Pantic [17] attempted to recognize the face expressions with the use of point's localization to detect the 20-face points. After tracking these points and applying hidden markov model (HMM), SVM, and GentleBoost to image sequence. [18] Temporal graphical model, like conditional random field (CRF) or HMM, which used for facial expression recognitions failed to discover various emotions. A hidden conditional ordinal random field (CORF) used to cover these issues by achieving both dynamic recognitions of various emotions and facial expression intensity estimation.

Verburg and Menkovski [14] presented a technique to recognize micro-expression by the use of optical flow features with RNN network to programme the temporal changes in face depend on histograms of oriented optical flow (HOOF) features extracting. The RNN networks spot slight intervals that are possible to have intersected facial micro movements. SAMM dataset used to train the RNN network, and by the use of leave-one subject outcross validation protocol. Their results display that with partial of the related micro movements there are 1569-false positives.

Sang et. al. [15], present a 3D-Inception ResNets networks (3DIRs), to recognize and process image sequence, which extend the 2D-Inception ResNets components. The added dimension extracts the spatial-relationships among the sequences of frames, which may results in a large of features map. A long short-term memory (LSTM) presented to follow the 3DIRs networks, which obtains the temporal-relationships to uses for sequence categorizing.

Minaee and Abdolrashidi [8], present a deep learning system established on convolutional network, which concentrates on the critical portions of faces with the use of visualization procedures based on the classifier's outputs. Their result presented that transformed emotions involve sensitive change in face portions. The present method realizes major advances over prior methods on numerous databases, containing FER2013, CK+, FERG, and JAFFE datasets.

Gu et. al. [9], present RNN network to track and join the facial features estimation in video frames. Their result showed that on the face-landmarks localization and heads poses estimations of video, performs on frames wise methods and Bayesian filtering to produce a large-scale datasets.

ISSN: 2005-4238 IJAST
Copyright ⓒ 2020 SERSC

3. Recurrent Neural Network Concept

RNNs are sorts of artificial neural network that enlarges extra weights to the networks to generate rotations in the networks graph in an effort to keep up an internal states. The possibilities of adding states to neural network, which may allow the network to be able to plainly learn and exploit setting in sequences predictions problems, such as the problem of an order or temporal component [13].

An improved way to think through this is the training of sets covers instances with a set of inputs for the current existing of training example. This is the traditional multi-layered perceptron, conventional method [17]. When the training example is complement with sets of inputs from the earlier example. This is a recurrent neural network, unconventional method [18].

As with all feed forward networks models, the concerns are how to link the input-layer to the output-layer, consist of feedback activation to train the construct to be joined [19].

RNNs learn temporal features with representing input-sequences to sequences of hidden states, additionally representing the hidden state to output-sequences [20]. RNNs have presented inspiring performance on numerous tasks. Training on long-term sequences is not an easy task, mainly due to exploding and vanishing problems [21]. LSTM solved these problems by providing a memory, which can remember/forget the data from earlier states for long period of time [22]. An LSTM unit has three gates:

1) Inputs gate (i)
2) Forget gate (f)
3) Output gate (o)

The block diagram of the LSTM architecture illustrated in Figure (1), [23].

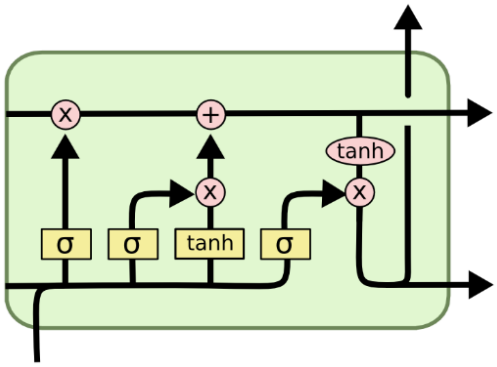


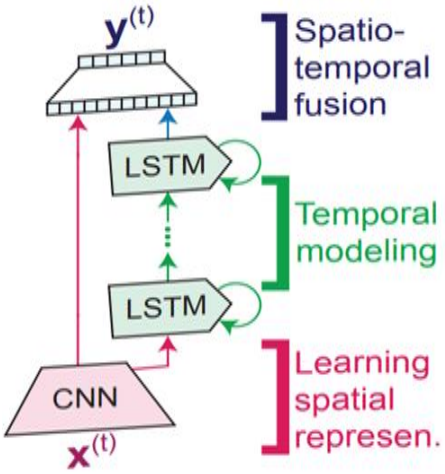
Figure (1): The LSTM Architecture.

4. The Proposed Model

LSTM layers change the hidden layers of classic RNNs. Each layer controls three elements (input-output-forget) gates for (read-reset-write) processes and contains one or more memory cells. The gates control the processes of the cells.

The proposed combination of CNN-LSTM is hybrid structure to extract spatial and temporal data of pain expression recognition. The hybrid pain recognition structure is illustrated in Figure (2) to extract face features by a 3-layers CNN 2-fully-connected layers.

Copyright ⓒ 2020 SERSC



模型完成了三层特征学习：
CNN：控件特征学习
LSTM：时域特征学习
CNN+LSTM：时域与空域特征融合

Figure (2): The Hybrid Combination of CNN-LSTM Architecture.

After, cropping the face and extracting the features from last fc layer of CNN as an output, this output used as input to a LSTM, which use the input, forget, and output gates, that learn the long-term dependencies and solving the vanishing or exploding gradient problem, Figure (3).

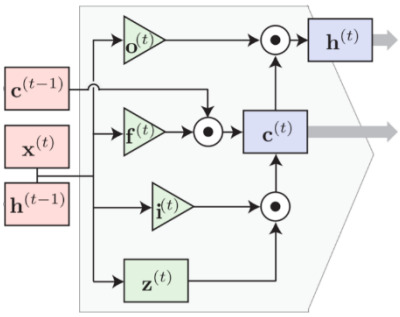


Figure (3): The LSTM Architecture.

If c_t is the total input at (t, t – 1) times, the input gate starts vector are calculated as:

$$i_t = \sigma\left(W_{yi}y_t + W_{hi}h_{t-1} + W_{si}s_{t-1} + b_i\right) \quad (1)$$

Where σ denotes sigmoid function, W is the weights matrix, y_t is the input at t, h_{t-1} is the hidden state vectors of the earlier times, and bi represents the input bias vector. Furthermore, the forget gate calculate as:

$$f_t = \sigma\left(W_{yf}y_t + W_{hf}h_{t-1} + W_{sf}s_{t-1} + b_f\right) \quad (2)$$

The memory cell value c_t is to be calculate as:

$$c_t = i_t tanh\left(W_{ys}y_t + W_{hs}h_{t-1} + b_s\right) + f_t \cdot s_{t-1} \quad (3)$$

The activations values of the output gates are to be calculate as:

$$o_t = \sigma\left(W_{yo}y_t + W_{ho}h_{t-1} + W_{so}s_{t-1} + b_o\right) \quad (4)$$

Memory cell output is:

$$h_t = o_t \cdot tanh(c_t) \quad (5)$$

Copyright ⓒ 2020 SERSC

Using bidirectional RNN-LSTM operational design, the input images/frames are processed by two different hidden layers to obtain extra features; h_{tf} , h_{tb} are the forward and backward activations values of the hidden layer, and the output layer update as:

$$z_t = W_{fz}h_t^f + W_{bz}h_t^b + b_y \quad (6)$$

Whereby the hidden bias-vectors, and W_{fy} and W_{by} is are the forward and backward weight matrices.

LSTM used to allocates the locations of expressions. LSTM offers points estimations that are more particular, compared with normal neural network.

Labels are expected prediction sequences-wise, with sequences input of n frames $f_i \in \{f_1, ..., f_n\}$, the target expectation is the pain expression in the f_n frame. Thus, training dataset is the data contain the (n-1) frames that will be used to estimate the pain expression in the current frame.

5 The Experimental Dataset and Result

5.1 The Experimental Dataset

The UNBC-McMaster Shoulder Pain Expression Archive (UNBC-MSPEA) dataset composed by researchers at McMaster University and University of Northern British Columbia. And its details as follows:

- 200 video sequences.
- 129 contributors (63 males and 66 females).
- 48403 FACS coded frames.

The patients in this dataset suffering from shoulders pain while performing a sequence of passive and active range of motions trials.

Figure (4), displays examples of the UNBC-MSPEA database.

Figure (4): Samples of Pain and Non-Pain Frames from the UNBC-MSPEA Dataset.

The proposed system generated various sets from the UNBC-McMaster datasets by modifying the frames resolution (by 25% and 50%) with the use of SR-algorithm [24]. The final estimate values for an image positions(x,y) are obtained as softmax form:

$$\mathrm{Y(x,y)} = \sum_{i=1,..,N} w_i(x,y)Z(x,y,i) \quad (7)$$

where

Copyright ⓒ 2020 SERSC

$$w_i(x,y) = exp\left(-\frac{|d_j(x,y)|}{\sigma_C}\right) / \left[\sum_{i=1,..,N} \exp\left(-\frac{|d_j(x,y)|}{\sigma_C}\right)\right] \quad (8)$$

and Z is the generated image.

5.2 Analysis of Experimental Results

The result of UNBC-McMaster Shoulder Pain dataset on the proposed hybrid deep learning architecture. Based on the FACS code, each video frame has pain different index between (0-16), there are two classes either no-pain (if the index less than 6), or pain (if the index more than 6). Two datasets were generated from the original UNBC-McMaster dataset. The LSTM networks created through 2-layers, and the performance assessed with the protocol of leave one-subject-out cross validation.

Table 1 and Figures 5 and 6, demonstration the results of the proposed system against the three datasets (original dataset(D1) and two generated datasets (D2 and D3)).

Table (1): The Accuracy Average Results.

Semantic Ground Truth	Pain_Index	D1	D2	D3
No_Pain	<6	60.1%	67.51%	60.28%
Strong_Pain	>6	74.2%	69.6%	86.34%
F1_score		0.69	0.68	0.72
Average	0-16	67.15%	68.56%	73.31%

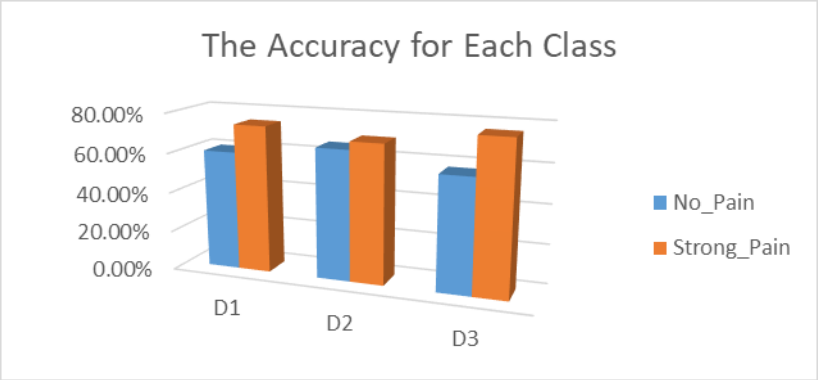


Figure (5): The Accuracy for each Class.

Copyright ⓒ 2020 SERSC

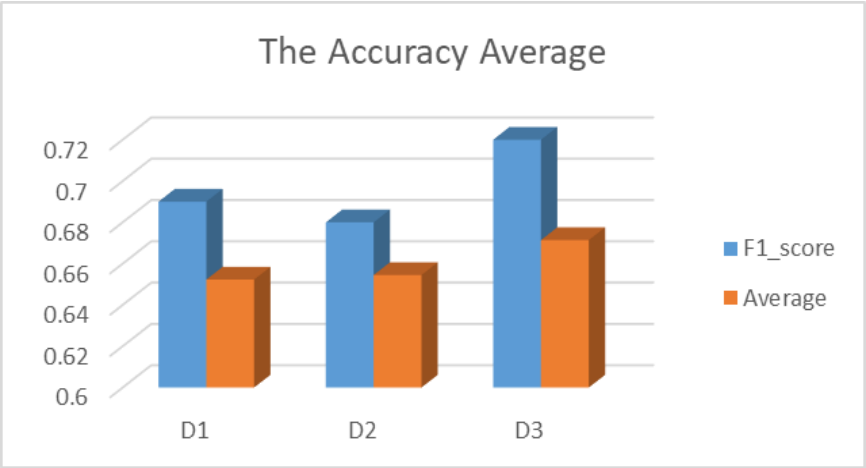


Figure (6): The Accuracy Average for each Dataset.

By discarding all the samples of weak pain and considered as no pain, the proposed system performance is slightly better on detecting the strong pain due to the huge changes on the facial expressions compared to the no pain class.

To check the effectiveness of using spatial and temporal data for classification results, the results of using a linear classifiers CNN structure without LSTM architecture, using spatial data only, with a comparison to the result from the temporal data only by the use of LSTM only in Table 2.

From the results in Table 2, the temporal data increases the average of predictions accuracy by 20%, and the accuracy differences mean that the spatial features are not sufficient for detecting the pain expressions in most cases. Therefore, by using the temporal features together with spatial features lead to improve the accuracy average in most cases.

For different subjects, different performance result due to the fast changes of facial expressions and the lighter pattern form among successive frames.

Table 2: Comparison of accuracy average results of CNN with the results of LSTM performances on the original dataset. The CNN depend on the data from single frame, however the LSTM depends into the differences of the temporal data of successive frames, thus the LSTM improves the accuracy estimation average by 20%.

Table (2): Comparison Of CNN With LSTM Accuracy Average For The Original 14 Subjects

Subject	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**AVG**
CNN	45.4	54.8	56.5	55.2	54.8	47.0	33.8	51.8	54.9	48.5	54.5	55.8	56.4	34.7	**50.3**
LSTM	67.0	70.5	72.0	74.6	91.5	57.0	37.0	49.0	90.0	74.4	91.0	75.0	74.5	69.5	**70.9**

6. Conclusions

By the use of modified architecture of CNN and LSTM, and the use of a both spatial and temporal data combination to get a high level features for pain recognition. The proposed technique assessed on UNBC_McMaster dataset with two new generated datasets from the original one by modifying the frames resolution (by 25% and 50%) with

Copyright ⓒ 2020 SERSC

the use of SR-algorithm [24]. The pain recognition performances concluded from the experimental results obtain worthy recognitions.

The experiment results showed also that including deep temporal data with the modified architecture increases the system abilities in recognitions between different kinds of expressions and this leads to a performances improvement of the proposed architecture. The results of the proposed process success to detect the pain expressions for the video sequence with an accuracy of 73.31%.

References

[1] Vallabh, Pranesh & Malekian, Reza, "Fall detection monitoring systems: a comprehensive review". Journal of Ambient Intelligence and Humanized Computing. 10.1007/s12652-017-0592-3, (2017).
[2] Zhang, T., Wang, J., Xu, L., Liu, P., "Fall detection by wearable sensor and one-class svm algorithm". Intelligent computing in signal processing and pattern recognition, 858–863 (2006).
[3] James R. Hagler, Alison L. Thompson, Melissa A. Stefanek, Scott A. Machtley, J. Insect Sci., "Use of Body-Mounted Cameras to Enhance Data Collection: An Evaluation of Two Arthropod Sampling Techniques", DOI: 10.1093/jisesa/iey033. (2018).
[4] Konstantinos Spaniolas, Julius D. Cheng, Mark L. Gestring, Ayodele Sangosanya, Nicole A. Stassen, Paul E. Bankey. **"**Ground Level Falls Are Associated with Significant Mortality in Elderly Patients". The Journal of Trauma: Injury, Infection, and Critical Care, 69 (4): 821 (2010).
[5] De la Concepcion, M. A. A., Morillo, L.M.S., Garcıa, J.A. A., Gonzalez, L., "Mobile activity recognition and fall detection system for elderly people using ameva algorithm". Pervasive and Mobile Computing 34, 3–13, (2017).
[6] Pannurat, N., Thiemjarus, S., Nantajeewarawat, E., "A hybrid temporal reasoning framework for fall monitoring". IEEE Sens. 17, 1749–1759 (2017).
[7] Yang, L., Ren, Y., Zhang, W., "3D depth image analysis for indoor fall detection of elderly people". Digital Communications and Networks 2(1), 24–34 (2016).
[8] Miaou, S.-G., Sung, P.-H., Huang, C.-Y., "A customized human fall detection system using omni-camera images and personal information". In: Distributed Diagnosis and Home Healthcare, D2H2. 1st Transdisciplinary Conference On, pp. 39–42 (2006).
[9] Haval A. Ahmed, Tarik A. Rashid, Ahmed T. Sadiq, "Face Behavior Recognition Through Support Vector Machines", (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 1, (2016).
[10] Liu, L., Ouyang, W., Wang, X., "Deep Learning for Generic Object Detection: A Survey". Int. J. Comput. Vis 128, 261–318, https://doi.org/10.1007/s11263-019-01247-4, (2020).
[11] Wafaa M. Salih, Ibraheem Nadher, Ahmed T. Sadiq, "Deep Learning for Face Expressions Detection: Enhanced Recurrent Neural Network with Long Short Term Memory". In: Applied Computing to Support Industry: Innovation and Technology. ACRIT 2019. Vol 1174. Springer, Cham, Iraq, (2020).
[12] Ji S, Xu W, Yang M, Yu K. 3D convolutional neural networks for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. (2013) 35(1):221–231.
[13] Nizam, Y., Mohd, M.N.H., Jamil, M.M.A., "Human fall detection from depth images using position and velocity of subject". Procedia Computer Science 105, (2017), 131–137.
[14] Rimminen, H., Lindstr¨om, J., Linnavuo, M., Sepponen, R., "Detection of falls among the elderly by a floor sensor using the electric near field". IEEE Transactions on Information Technology in Biomedicine 14(6), 1475–1476, (2010).

Copyright ⓒ 2020 SERSC

[15] Simonyan, K.; Zisserman, A., "Two-stream convolutional networks for action recognition in videos". In Proceedings of the Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems, Montreal, Canada, 8–13 December (2014).
[16] Salah Sleibi Al-Rawi, Ahmed T. Sadiq, Wasan M. Alaluosi, "Feature Extraction of Human Facail Expressions Using Haar Wavelet and Neural network", Iraqi Journal of Science, Vol. 57, No.2C, (2016), pp:1558-1565.
[17] Karpathy, Andrej, Joulin, Armand, and Fei-Fei, Li. "Deep fragment embeddings for bidirectional image sentence mapping". In arXiv:1406.5679, (2013).
[18] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. "Learning spatiotemporal features with 3D convolutional networks". In Proceedings-IEEE International Conference on Computer Vision, ICCV, (2015) 11-18-December, pp. 4489-4497.
[19] Wafaa M. Salih Abedi. "Unconsciousness detection supervision system using faster RCNN architecture", Proceedings of the 2nd International Conference on Future Networks and Distributed Systems - ICFNDS '18, Amman, Jorden, (2018).
[20] Peng X., Schmid C., "Multi-region Two-Stream R-CNN for Action Detection". In: Computer Vision – ECCV 2016. Lecture Notes in Computer Science, vol 9908. Springer, Cham, (2016).
[21] Catalin Ionescu, Dragos Papava, Vlad Olaru and Cristian Sminchisescu, "Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, No. 7, July (2014).
[22] Catalin Ionescu, Fuxin Li and Cristian Sminchisescu, "Latent Structured Models for Human Pose Estimation", International Conference on Computer Vision, (2011).
[23] Adhikari, Kripesh, Hamid Bouchachia, and Hammadi Nait-Charif, "Activity recognition for indoor fall detection using convolutional neural network", Machine Vision Applications (MVA), Fifteenth IAPR International Conference on. IEEE, (2017).

Copyright ⓒ 2020 SERSC

View publication stats