Recognition and Visualization of Facial Expression and Emotion in Healthcare

Hayette Hadjar[1], Thoralf Reis[1], Marco X. Bornschlegl[1], Felix C. Engel[2], Paul Mc Kevitt[3], and Matthias L. Hemmje[2]

[1] University of Hagen, Faculty of Mathematics and Computer Science, 58097 Hagen, Germany
{hayette.hadjar, thoralf.reis, marco-xaver.bornschlegl}@fernuni-hagen.de

[2] Research institute for Telecommunication and Cooperation, FTK, Dortmund, Germany
{fengel, mhemmje}@ftk.de

[3] Ulster University, Derry/Londonderry, Northern Ireland
p.mckevitt@ulster.ac.uk

Abstract. To make the SenseCare KM-EP system more useful and smart, we integrated emotion recognition from facial expression. People with dementia have capricious feelings; the target of this paper is measuring and predicting these facial expressions. Analysis of data from emotional monitoring of dementia patients at home or during medical treatment will help healthcare professionals to judge the behavior of people with dementia in an improved and more informed way. In relation to the research project, SenseCare, this paper describes methods of video analysis focusing on facial expression and visualization of emotions, in order to implement an "Emotional Monitoring" web tool, which facilitates recognition and visualization of facial expression, in order to raise the quality of therapy. In this study, we detail the conceptual design of each process of the proposed system, and we describe our methods chosen for the implementation of the prototype using *face-api.js* and *tensorflow.js* for detection and recognition of facial expression and the *PAD space* model for 3D visualization of emotions.

Keywords: Emotion Recognition, Facial Expression Analysis, Emotion Visualization, Emotion Monitoring, Convolutional Neural Networks (CNNs), Affective Computing.

1 Introduction

Emotion analysis is important, because emotions penetrate many aspects of our lives, by informing the decisions we make and how we choose to communicate our thoughts to others and to ourselves. The data obtained can facilitate the diagnosis of emotional needs related to anxiety, depression or other kinds of mental illnesses. Healthcare for people with dementia living in their homes has become essential and health professionals need more information on patient behavior to prevent deteriora-

tion of their health. It is very important to intervene and predict this deterioration before it happens, avoiding the worst outcomes. In this context, the work reported here monitors the emotional state of the patient at home by recording videos, and analyzing video frames for facial recognition content and providing temporal trends and visualizations. The therapist or healthcare professional can intervene in the case of anxiety or depression for a dementia patient evident in the results before their mental state deteriorates further. Hence, primary care professionals will have an improved overview of the emotional wellbeing of patients through SenseCare [1]. SenseCare integrates data streams from multiple sensors and fuses data from these streams to provide a global assessment that includes objective levels of emotional insight, well-being, and cognitive state. There is potential to integrate this holistic assessment data into multiple applications across connected healthcare and various other inter-related and independent domains. SenseCare is thematically aligned with the current EU Horizon 2020 themes of "Internet of Things", "Connected Health", "Robotics" (including emotional robotics) and the "Human Brain Project". SenseCare has identified three application scenarios: (i) Assisted Living Scenario, (ii) Emotional Monitoring Scenario, (iii) Shared care giving Scenario. This work is focused on the second application scenario, *Emotional Monitoring* which deals with the emotional monitoring of people with dementia during medical treatment. Healthcare professionals will be better informed about the behavior of patient, by using SenseCare in order to raise the quality of therapy. Processing of voluminous data streams from video recordings, on the basis of the recently introduced Information Visualization for Big Data (IVIS4BigData) model [2] elaborates data stream types addressed by our visualization approach.

The visualization of emotions can be applied to Psychiatry and Neurology and relates to the diagnosis of people suffering from mental health problems such as e.g., dementia and depression. Human emotions are hypothetical constructs based on physiological and psychological data. The aim of the work reported here proposes to find solutions for the following points:

- The representation of emotions.
- The detection and recognition of emotion.
- The visualization of people's emotional states from video recordings, real-time and offline video analysis processing, and how users will comprehend the results.
- The visualization and exploration of emotion dynamically over time.

The remainder of this paper is organized as follows. Section 1 presents the state of the art of Convolutional Neural Networks (CNNs), facial Expression Emotion detection from video recordings, and visualization of dynamic emotion (over time). In Section 2 we detail conceptual designs of modeling of solutions for recognition and visualization of emotion. Section 3 describes our chosen methods for implementation of the prototype for recognition and visualization of facial expression and emotion, and finally we conclude in section 4 also discussing future work.

2 State of the Art

3 Methodology

Here we discuss methodology for operation of the SenseCare recognition and visualization of facial expression and emotion.

3.1 Video recognition process

Classifying video presents unique challenges for machine learning models. Video has the added property of temporal features in addition to the spatial features present in 2D images. An overview of the video recognition process using CNNs is shown in Fig. 4:

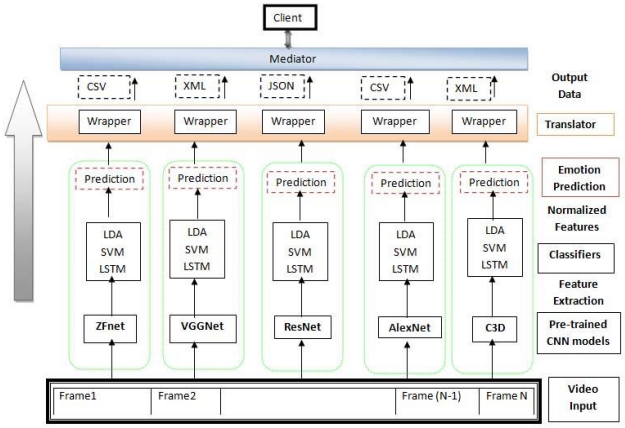


Fig. 4 Global Overview of Video Recognition Process using CNNs

As shown in Fig. 4, the video frames are split into individual frames and on each frame, face detection is executed. Each individual frame is fed back into the system for face feature detection. For *feature extraction*, in the first learning stage *video input* is pre-trained with CNN models such as C3D, AlexNet, ResNet, VGGNet or ZFnet. In the second learning stage on *training linear classifiers*:

Following feature extraction, there is a binary classification task for each frame employing:

- Long short-term memory (LSTM) is a particular type of Recurrent Neural Network (RNN) which performs better than the standard version [15]. The goal of LSTMs is to capture long-distance dependencies in a sequence, such as the context words.
- A Softmax classifier or a Support Vector Machine (SVM) classifier [20] employed to capture the emotion-specific information. The fusion of CNN and SVM classifiers provides better results compared to individual classifier performance.
- Linear discriminant analysis (LDA): a method employed recognition machine learning for face image to separate two or more classes of objects or events, for dimensionality reduction before later classification.

Emotion prediction: at the conclusion of processing, linear classifiers automatically predict emotional states by referring to training sets for *anger, disgust, fear, happi-*

ness, *sadness*, and *surprise*. Results are stored in data files in .CSV or .XML format, and a mediator provides output to the client.

3.2 Emotion Visualization Process

An overview of the emotion visualization process is detailed in Fig. 5.

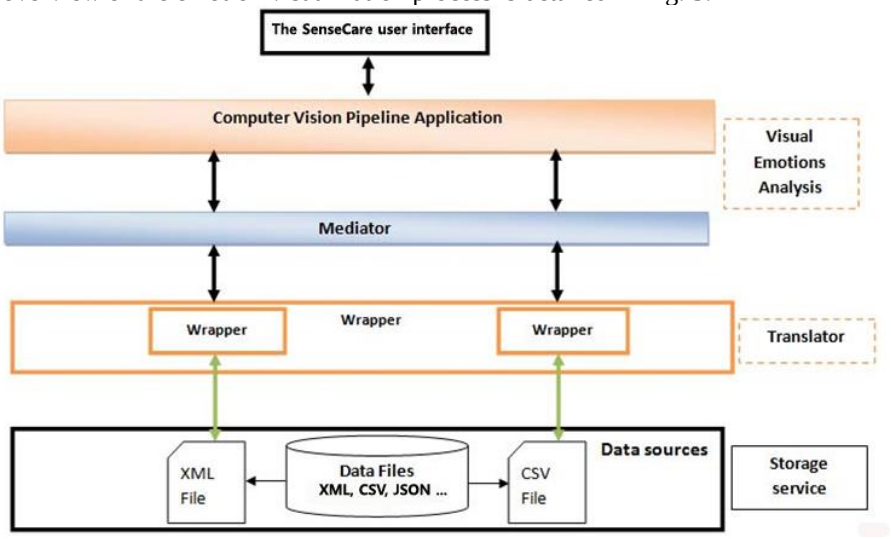


Fig. 5 Overview of Emotion visualization Process

The SenseCare user interface sends requests to the computer vision pipeline application. Results from video recognition are stored in data files in XML, CSV, or JSON format. The visualization prototype uses a wrapper to select data, and a web server in the mediator sends responding customized graphs of emotions as output results to the SenseCare user interface.

4 Implementation methods

For the implementation of the *SenseCare Emotional Monitoring* Scenario prototype, each operation requires specific hardware and software to succeed. Fig. 6 illustrates the workflow of real-time and offline facial expression video analysis. Real-time video analysis uses video streamed from the webcam as data input. Offline video analysis uses us input pre-recorded video files. The same processing operations will be completed in both processes using pre-trained CNN models for feature extraction and classifiers for emotion prediction.

Offline video analysis can quickly analyze multiple video files in parallel and provide useful alerts with snapshots to browse a large set of video data. Video Streams can continuously capture video data from the webcam and store terabytes of data per hour from hundreds of thousands of sources.

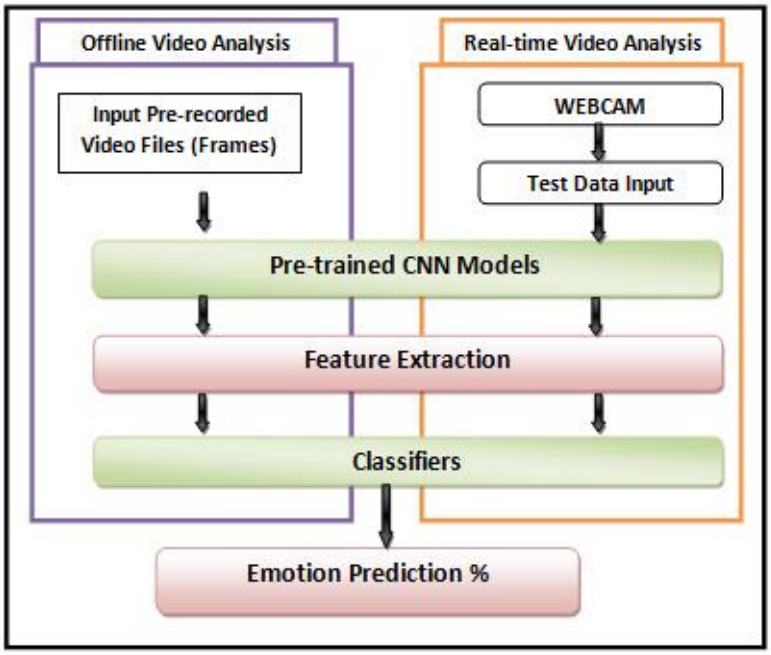


Fig. 6 The workflow of real-time and offline facial expression video analysis

4.1 Real-time video streaming

In Fig. 7 the architecture for remote streams video recovery is shown.

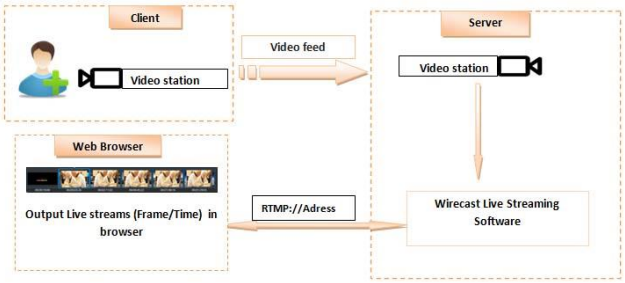


Fig. 7 SenseCare videoconferencing client/server relationship

The proposed solution employs two video stations (see Fig. 8), one in the client part (e.g., patient at home), and the second in the server part server part (e.g., at hospital). The Video stations consist of input and output ports to transmit video feeds. The external capture card plugs into the server using the appropriate Thunderbolt connection between Camera and Wirecast Live Streaming Software, to receive improved image quality. The RTMP (Real Time Messaging Protocol) address is added (e.g., public IP set) as a source for broadcast in the Wirecast Live Streaming software. The end-user must download and install the Flash browser plugin in order to playback

audio and video streamed by RTMP in a Web browser. This videoconferencing system architecture is employed in the WebTV of CERIST (Research Centre on Scientific and Technical Information) [21].

Fig. 8 Remote streaming video station at CERIST [21]

4.2 Real-time video analysis

The basic software components for real-time analysis of video frames taken from live video streams are as follows:

- Acquire frames from the video source
- Select the frames to be analyzed
- Submit these frames to the API
- Each result of the analysis returned is consumed by the API call

4.3 Detection and recognition of emotion

Essential steps in Face Recognition processing are summarized as follows:

1. *Face Detection*: Locate one or more faces in the image and mark with a bounding box.
2. *Face Alignment*: Normalize the face for consistency with the database, such as e.g., geometry and photometrics.
3. *Feature Extraction*: Extract features from the face that can be employed in the face recognition task (4.).
4. *Face Recognition*: Perform matching of the face against one or more known faces in a prepared database.

Face detection is an essential step in emotion recognition systems. It is the first step in our system to define a region of interest for feature extraction.

We have chosen *Face-api.js* [22] and *Tensorflow.js* [23] for video processing. *Face-api.js* is a JavaScript API for face recognition in the browser with *Tensorflow.js*. *TensorFlow.js* is a library for building and executing machine learning algorithms in JavaScript. *TensorFlow.js models* run in a web browser and in the *Node.js* environment. *TensorFlow.js* has empowered a new set of developers from the extensive JavaScript community to build and deploy machine learning models and enabled new classes of on-device computation [24]. For face recognition, a *ResNet-34* like architecture is applied to calculate a face descriptor, a feature vector with 128 values.

4.3.1 Face Detection, Landmark Detection and Alignment

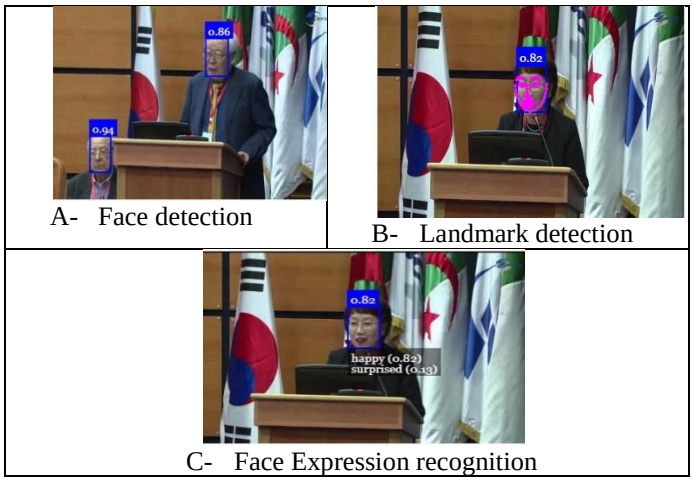
Face-api.js solely implements an SSD (Single Shot Multibox Detector) MobileNets v1 based CNN for face detection. MobileNets relies on a streamlined structure that uses detachable deep convolution to create deep and light neural networks.

This app also implements an optimized *Tiny Face Detector*, basically an even tinier version of *Tiny Yolo v2* [25] utilizing depthwise separable convolutions instead of regular convolutions, and finally also employed MTCNN (Multi-task Cascaded Convolutional Neural Network) [26]. The neural networks return the bounding boxes of each face, with their corresponding scores. The default model *face_landmark_68_model* and the tiny model *face_landmark_68_tiny_model* return the 68 point face landmarks of a given face image.

4.3.2 Face Expression Recognition

Face images can be extracted and aligned in the face recognition network, which is based on a *ResNet-34* type structure and basically matches the structure applied in dlib's face recognition model [27]. *Face-api.js* is prepared to evaluate faces detected and allocates a score (from 0 to 1) for each expression, *neutral*, *happy*, *sad*, *angry*, *fearful*, *disgusted*, *surprised*, in real-time with a webcam. The facial expression recognition model performs with reasonable accuracy. The model has a size of roughly 310kb and implements several CNNs. It was trained on a variety of images from open access public data sets as well as images extracted from the web. The API employs a *Euclidean distance* classifier to find best matches in */data/faces.json* [28].

The results are optimized for the web and for mobile devices. We have used videos from CERIST-WebTV for testing the API (see Fig. 9).



A- Face detection

B- Landmark detection

C- Face Expression recognition

Fig. 9 Demonstration of Face detection, Landmark detection, Face Expression Recognition.

In the case of real-time video, SenseCare selects a frame from the video and analyses the images every 500 ms, so the following step is how to return the results for a particular image in JSON format.

4.3.3 Confusion Matrix

A confusion matrix is employed in evaluating the correctness of a classification model. In classification problems, ‘accuracy’ refers to the number of correct predictions made by the predictive model over the rest of the predictions. The four public

face expression databases are CK+, Oulu-CASIA, TFD, and SFEW [30]. CK+ (Extended Cohn-Kanade dataset) consists of 529 videos from 123 subjects, 327 of them annotated with eight expression labels.

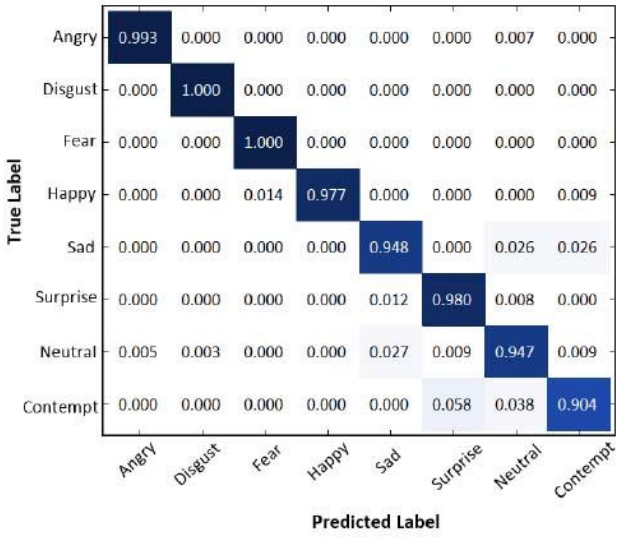


Fig. 10 Confusion Matrix of CK+ for the Eight Classes problem. (darker color = higher accuracy) [16].

Experiential results of the confusion matrix for face expression recognition in videos are shown in Fig. 10.

'Accuracy' refers to the number of correct predictions made by the predictive model over the rest of the predictions. Accuracy is used when the target variable classes in the data are nearly balanced, but is not used if the target variables in the data are the majority of one class.

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})}$$

TR=True Positive, TN=True Negative, FP= False Positive, FN= False Negative.

4.4 Emotion Visualization

There are two ways of representing affective states in video content: (1) discrete affective categories, and (2) continuous affective dimensions.

4.4.1 Implementing the Affective Model

Dimensional models of emotion attempt to visualize human emotions by determining where they are located in 2 or 3 dimensions. Most dimensional models include *valence* and *arousal* or *intensity* dimensions. For example, The *Circumplex model* of affect [30] defines two dimensions, *pleasure* and *arousal*, whilst the PAD (Pleasure-Arousal-Dominance) [31] emotional state model uses 3. *PAD space*: Pleasure-Arousal-Dominance is a psychological model proposed by Albert Mehrabian. The PAD model is employed in describing and measuring emotional states. It's one of the

3D models that that has gained popularity in affective computing. The PAD model allows us to differentiate anger (positive dominance) from fear (negative dominance).

As shown in Fig.11, dimensions (PAD-axes) are:

- Pleasure / Valence (P)
- Arousal (A)
- Dominance (D)

DES denotes Default Emotional State, *Ei*, Emotions, and *ES(t)*, Emotional State.

The following labels describe the resulting octants of the PAD model:

(+P+A+D) *Exuberant*, (-P-A-D) *Bored*, (+P-A+D) *Relaxed*, (-P+A-D) *Anxious*, (+P+A-D) *Dependent*, (-P-A+D) *Disdainful*, (+P-A-D) *Docile*, (-P+A+D) *Hostile*.

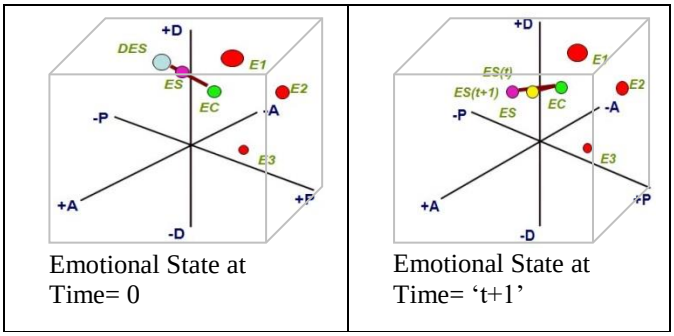


Fig. 11 Exploration of emotions in PAD model space [32]

Intensity of Emotional State (ESi) is defined as follows:

ESi = ||ES||, if ESi ∈ [0.0, 0.57], *Slightly*

ESi ∈ [0.57, 1.015], *Moderate*

ESi ∈ [1.015, 1.73], *Highly*

Emotional data is mapped in PAD space and takes the following form:

Visualization space = {PAD positions (XYZ-axes), Val, color of emotion}.

Val=% emotion prediction.

Building interactive visualizations using WebVR and NodeJS technologies [33] allows synchronous or real-time communication in the web application.

Fig.12 shows how the Websocket operates in SenseCare, including user do action change position or rotation, socket emit action (socket emit), the server NodeJS received data, and sends (socket on) to update information for all users connected in this space.

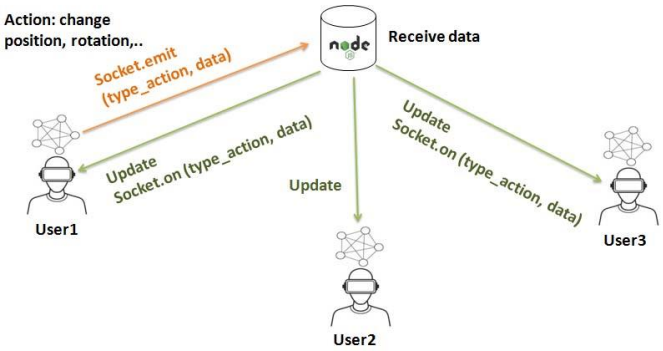


Fig. 12 Real-time collaborative data communication and visualization

Fig. 13 shows simulation of emotion dynamics in the PAD model space:

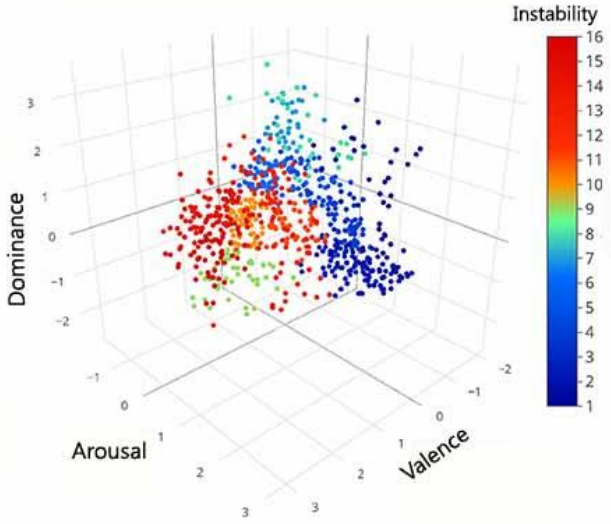
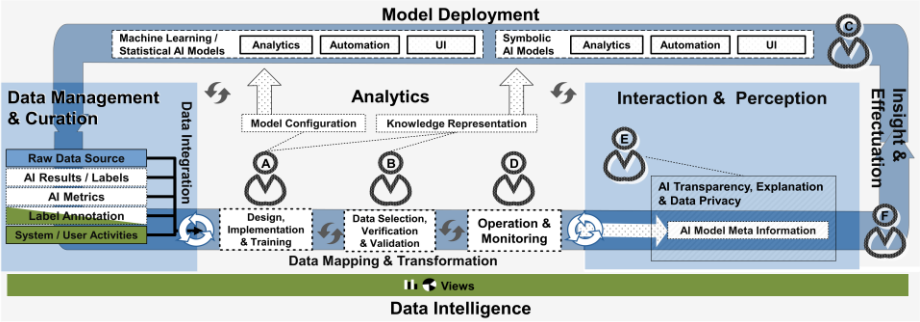


Fig. 13 Simulation of PAD annotation distributions of video facial expression analysis

4.5 AI2VIS4BigData model

Processing large data streams from video recordings (in real-time and offline) to be visualized fits with the AI2VIS4BigData reference model (see Fig. 14). AI2VIS4BigData model [34] is based on AI transparency, explanation, and data privacy. It is also based on the life cycle of the AIGO AI system [35] and the reference model IVIS4BigData [2] of Bornschlegl for the analysis and visualization of Big Data.



A = Model Designer, B = Domain Expert, C = Model Deployment Engineer, D = Model Operator (MLOps), E = Model Governance Officer, F = Model End User.

Fig. 14 AI2VIS4BigData: A reference model for AI supporting Big Data analysis [34].

All the elements of the AI2VIS4BigData model such as AI models, user stereotypes, and data management are relevant to our visualization approach.

5 Conclusion and future work

In order to develop major advances in emotion analysis, there must be adequate techniques for combining and analyzing complex signals. This research explores two architectures for the *Emotional Monitoring Scenario*, for processing video of a dementia patient recorded at home. Existing techniques for video face expression recognition employing CNNs and classifiers are detailed, including the exploration of emotional states using *PAD Space* model. Future work includes:

a- Finding solutions to manage and store (streaming & offline) video analysis results in the cloud (format JSON, CSV, XML, Databases).

b- Developing a web tool for visualization of emotional states and possibly including collaborative visualization.

c- Testing the SenseCare prototype in a live setting with live patients.

Future improvements will include new visual representations, views, and the collection of additional types of data such as eye-movement monitoring.

Our primary challenge is to develop a cloud-based affective computing system capable of processing and fusing multiple sensory data streams to provide cognitive and emotional intelligence for AI connected healthcare systems employing sensory and machine learning technologies, in order to augment patient well-being with more effective treatment across multiple medical domains.

Acknowledgements

This research has been developed in the context of the SenseCare project. SenseCare has received funding from the European Union's H2020 Programme under grant agreement No 690862. However, this paper reflects only the authors' views and the

European Commission is not responsible for any use that may be made of the information it contains.

References

1. Engel, F., Bond, R., Keary, A., Mulvenna, M., Walsh, P., Hiuru, Z., Kowohl, U.,Hemmje, M.L.: Sensecare: Towards an experimental platform for home-based, visualisation of emotional states of people with dementia. Computer Science, Springer, 2016 (Engel, et al., 2016).
2. M. X. Bornschlegl, K. Berwind,M. Kaufmann, F. C. Engel, P.Walsh, M. L. Hemmje,and R. Riestra, "IVIS4BigData: A reference model for advanced visual interfaces supporting big data analysis in virtual research environments", Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10084 LNCS, pp. 1-18, 2016.
3. Goleman, D. (1995). Emotional intelligence. Bantam Books, Inc.
4. R. R. Bond, H. Zheng, H. Wang, M. D. Mulvenna, P. McAllister, K. Delaney , P. Walsh, A. Keary, R. Riestra and S. Guaylupo, "SenseCare: using affective computing to manage and care for the emotional wellbeing of older people," in eHealth 360°, vol. 181, K. Giokas, B.Laszlo and F. Hopfgartner, Eds., Springer, 2017, pp. 352-356.
5. Machine Intelligence and Signal Processing, Ebook, Proceedings of International Conference, Springer, Singapore, MISP 2019, ISBN 978-981-13-0923-6.
6. R. A. Minhas, A. Javed, A. Irtaza, M. T. Mahmood, and Y. B. Joo, ''Shot classification of field sports videos using alexnet convolutional neural network,'' Appl. Sci., vol. 9, no. 3, p. 483, 2019.
7. K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In ICLR, 2015.
8. Brownlee, J. Deep Learning for Computer Vision: Image Classification, Object Detection, and Face Recognition in Python; Machine Learning Mastery, Vermont, Australia, 2019.
9. LIM, Y. K., LIAO, Z., PETRIDIS, S., AND PANTIC, M. Transfer learning for action unit recognition. CoRR abs/1807.07556 (2018).
10. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 770–778, 2016.
11. Byoungjun Kim and Joonwhoan Lee, "A Deep-Learning Based Model for EmotionalEvaluation of Video Clips",International Journal of Fuzzy Logic and Intelligent Systems. Vol. 18, No. 4, December 2018, pp. 245-253.
12. Turabzadeh, S.; Meng, H.; Swash, R.M.; Pleva, M.; Juhar, J. Facial Expression Emotion Detection for Real-Time Embedded Systems. Technologies 2018, 6, 17.
13. Bahreini, K., van der Vegt, W. & Westera, W. A fuzzy logic approach to reliable real-time recognition of facial emotions. Multimed Tools Appl 78, 18943–18966 (2019).
14. Guérin-Dugué A, Roy RN, Kristensen E, Rivet B, Vercueil L and Tcherkassof A (2018) Temporal Dynamics of Natural Static Emotional Facial Expressions Decoding: A Study Using Event- and Eye Fixation-Related. Potentials. Front. Psychol. 9:1190.doi: 10.3389/fpsyg.2018.01190
15. "Long short-term memory". Neural Computation. 9 (8): 1735–1780. doi:10.1162/neco.1997.9.8.1735. PMID 9377276.
16. Ding,H , Zhou,SK, Chellappa, R:'FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition' In 12th IEEE International Conference on Automatic Face and Gesture Recognition, pp.118-126(2017)

17. Chitra Vasudevan, Concepts and Programming in PyTorch: A way to dive into the technicality, BPB Publications, 2018, ISBN 9388176057, 9789388176057
18. OpenCV (Open Source Computer Vision Library), link: https://opencv.org/ , (viewed 23 June 2020).
19. Adrian Rosebrock, Live video streaming over network with OpenCV and ImageZMQ , https://www.pyimagesearch.com/2019/04/15/live-video-streaming-over-network-with-opencv-and-imagezmq/, (viewed 23 June 2020).
20. Sreenivasa Rao Krothapalli, Shashidhar G. Koolagudi, “Emotion Recognition using Speech Features", 2013, Springer New York, doi.org/10.1007/978-1-4614-5143-3.
21. Research Centre on Scientific and Technical Information, link: http://www.cerist.dz, (viewed 23 June 2020).
22. Face-api.js, JavaScript API for face detection and face recognition in the browser and nodejs with tensorflow.js, link:https://github.com/justadudewhohacks/face-api.js/, (viewed 23 June 2020).
23. TensorFlow.js, JavaScript library for machine learning, link: https://www.tensorflow.org/js, (viewed 23 June 2020).
24. Daniel Smilkov, Nikhil Thorat, Yannick Assogba, Ann Yuan, Nick Kreeger, Ping Yu, Kangyi Zhang, Shanqing Cai, Eric Nielsen, David Soergel, et al. Tensorflow.js: Machine learning for the web and beyond. arXiv preprint arXiv:1901.05350, 2019.
25. Tiny YOLO v2 object detection with tensorflow.js, Link: https://github.com/justadudewhohacks/tfjs-tiny-yolov2 , (viewed 23 June 2020).
26. K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Processing Letters, 23(10):1499–1503,2016.
27. Dlib C++ Library, link: http://dlib.net/, (viewed 23 June 2020).
28. Realtime Face Recognition in the Browser, link: https://morioh.com/p/ddbc538212df, (viewed 23 June 2020).
29. H. Ding, S. K. Zhou and R. Chellappa, "FaceNet2ExpNet: Regularizing a Deep Face Recognition Net for Expression Recognition," 2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017), Washington, DC, 2017, pp. 118-126, doi: 10.1109/FG.2017.23.
30. Russell, James (1980). "A circumplex model of affect". Journal of Personality and Social Psychology. 39 (6): 1161–1178. doi:10.1037/h0077714.
31. Mehrabian, A. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. Current Psychology 14, 261–292 (1996).
32. D.D.L. Arellano Tavara, Visualization of Affect in Faces Based on Context Appraisal. Doctoral Thesis, University of Balearic Islands, Spain, 2012.
33. H.Hadjar, A.Meziane, R.Gherbi, I.Setitra, N.Aouaa. 2018. WebVR based Interactive Visualization of Open HealthData. In International conference on Web Studies (WS.2 2018), October 3–5, 2018, Paris, France. ACM, New York, NY, USA, 8 pages.
34. T. Reis, M. X. Bornschlegl, and M. L. Hemmje, \Towards a Reference Model for Artificial Intelligence Supporting Big Data Analysis," Proceedings of the 2020 International Conference on Data Science (ICDATA'20), 2020.
35. OECD, Artificial Intelligence in Society, 2019.
36. Keras implementation of residual networks, link: https://gist.github.com/mjdietzx/0cb95922aac14d446a6530f87b3a04ce , (viewed 23 June 2020).