FACE EMOTION RECOGNITION USING A CRISIS RELATED SMARTPHONE APP

Dragos Datcu, Leon Rothkrantz***

* InteliWatch, Delft, The Netherlands
** Delft University of Technology, Delft, The Netherlands
** Czech Technical University in Prague, Prague, Czech Republic
The Netherlands, Czech Republic

** Corresponding Author, e-mail: L.J.M.Rothkrantz@TUDelft.nl

Abstract: Especially during a natural crisis, stressful events may happen and people use social media tools to ask for help or assistance. Priority is given by rescue workers, to those people in high physical or psychological need. In case a selfie is attached to the message by the user, face analysis can be used to assess the emotional state of the user. In this paper we present a system to recognize emotions in facial expressions using Convolutional Neural Networks and AffectNet Database.

Key words: Facial Expression Recognition, Crisis Events, Mobile Systems, Convolutional Neural Networks, AffectNet Database.

1. INTRODUCTION

In this paper we report about the design of a system which enables emotional analysis of facial expressions. The model is based on Convolutional Neural networks and has been trained on a huge database of facial expressions recorded in the wild such as AffectNet database [1]. The usual databases of facial expressions are recorded in a laboratorial environment, which limits the recognition rate in real life applications. Selfies are usual frontal pictures, under good lighting conditions and absence of occlusion. This facilitates the analysis of facial pictures enormously.

Nowadays we can see that many people use their smart phone constantly to report about their experiences and observations. Usual these messages are processed by human receivers. During a natural crisis many messages are communicated. The messages provide information for the crisis team to update their crisis awareness. To analyse what is going on and if people are in need, instantly processing of the messages is necessary. This calls for automated systems to analyse these messages and search for emotional content of these messages. In

[2] we analysed verbal messages by using keyword spotting. In this paper we analyse attached facial expressions automatically to assess the emotional state of the sender. Fast processing is needed to localise persons in physical or psychological need and to plan evacuation or assistance.

Emotions play an important role in human communication. People are able to assess the psychological state of themselves and from other people and to communicate this information. In the same way appreciations of situations, actions and products can be evaluated using emotions. Emotions can be displayed in a verbal and nonverbal way using body language as posture, gestures and facial expressions. Ekman [3] was able to define 6 universal emotions happiness, sadness, anger, fear, surprise, disgust and contempt. To assess emotions automatically, we have to solve the problem of lighting, posture and occlusion. In case of selfies the user is able to control these parameters so he or she can influence the quality of the input video.

In daily life people use blended emotions of the basic emotions of different intensity. Different models have been developed of blended emotions. An alternative solution of the discrete emotion space is the representation of emotions using a continuous 2D space. The horizontal axis can be used to represent the positive or negative aspects (valence) and the vertical axis can be used to represent the active or passive aspects of emotions (arousal). Well-known is the circumplex model of affect of Russell [4] displayed in figure 1.

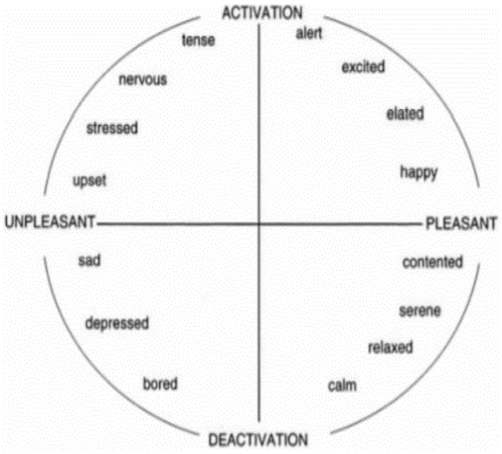

Fig. 1. A Graphical representation of the circumplex model of affect (Russell)

Our focus of interest are emotions shown during a natural hazard and are supposed to have a negative connotation. This are emotions in the left half of Russell's circumplex. It is common use to describe facial emotions by changing contours of the eyes, eyebrows and the mouth.

Currently, deep learning algorithms are used to analyse facial expressions. A huge database of recorded facial expression is needed and a long processing time to

train a system for automated assessment of facial expressions. We got access to huge databases of recorded facial expressions [1]. We developed a special facial model based on facial appearance technology. It may be expected that people show a lot of emotions during hazards especially at the peak of a hazard. To record facial expressions, we developed a special Crisis App (see figure 2, right) to send messages and attached photos to the network. We chose for local processing to overcome overloaded networks.

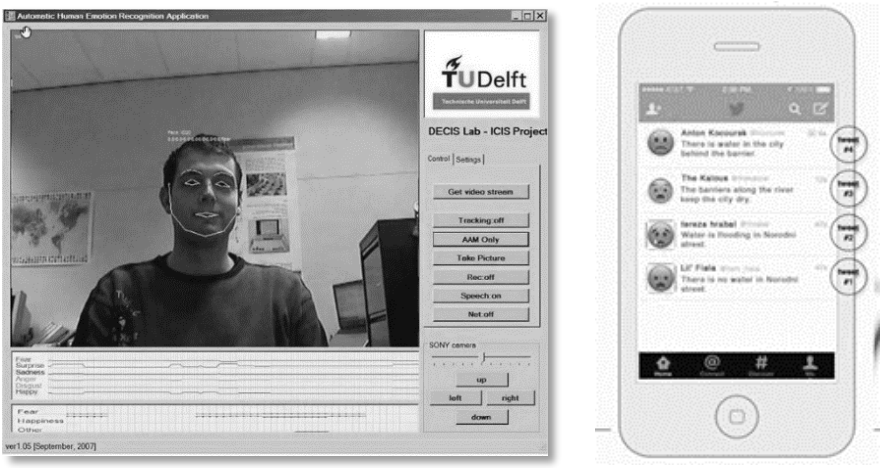

Fig. 2. The tool to analyse facial expressions (left) and the Crisis App (right)

The outline of the paper is as follows. In section 2 we discuss related works. In section 3 we present the architecture of the system as well as the underlying computer vision and deep learning algorithms. In section 4 we present and discuss on the findings of our experiments. Section 5 presents the conclusions of the study.

2. RELATED WORK

Social media play an important role in disaster management and providing help to people in need, a first problem is to differentiate between disaster-related messages and non-disaster-related messages. Many AI technologies have been used varying from keyword technology to high sophisticated machine learning technology. In [2] a straightforward approach based on keyword technology and rule-based was used to analyse tweets recorded during crisis events related to natural hazards. For every disaster a database of relevant keywords was created, used to detect the ongoing disaster.

To analyse and compare state of the art machine learning technology huge corpora of recorded tweets during a crisis are necessary. In [5], Wiegmann et al. introduced the Disaster Tweet Corpus 2020, a compilation of 123.166 tweets from 48 disasters covering 10 disaster types. A first problem was to annotate a great part of the tweets by human annotators for learning and testing. A second problem was

to handle the imbalance between different classes. Third to research cross-type setting problems, that is to say to train a model on one corpus and to test on another corpus. In total 162 models were used to analyse the data. Application of Convolution Neural Networks and Bert resulted in a good performance.

Svensson and Pesämaa [6] emphasize that emotional expressions become an important aspect of situational assessment and decision-making during emergency calls. For instance, on one way fear is proven to help emergency call operators as they execute emergency decisions. On the other way, fear seems to be more difficult to assess than anger and sadness. The expressed emotions of the callers and their perception by the professional personnel affect the decisions and impacts on the response time for getting help. In this context, supporting emergency operators better evaluate the emotion frame provides valuable support in properly handling emergencies. People report emergency incidents typically by voice and through their mobile phones. We develop a system which relies on a mobile app to process live video feed from the smartphone's camera and subjectively senses the caller's emotion during the emergency call. The emotions are perceived by sensing facial expressions while the phone is kept in hand and pointing to the caller's face. This relates to the situation when the caller sets the device on speaker more or with the headphones connected.

The group Knowledge Based Systems was for many years involved in the automated recognition of emotions in facial expressions, speech and body language. Datcu developed an automated system for analysing facial expression based on the Active Appearance Model [7]. The developed system was applied in crisis situation such as the flooding of the Vltava River at Prague [8]. A special Massive Open Online Course (MOOC) has been developed to train citizens of Prague in crisis awareness. In [9] experiments the developed Crisis App has been tested to improve mitigation models.

Since 2000, different Neural Network architectures were used to recognise emotional facial expressions. At start the available algorithms could only be applied to limited databases of recorded and annotated facial expressions. The application of Convolutional neural networks enables fast processing of huge databases. At start databases were composed of facial expressions recorded in a laboratory environment. But recently many new databases have been created composed of emotional facial expressions recorded in the wild [1]. This resulted in an exponential grow of papers using CNN and huge databases.

3. SYSTEM ARCHITECTURE

We developed an emotion-oriented system which assists remotely the emergency operator in better understanding and taking action while handling the emergency. The system reads the non-verbal emotion clues and informs the emergency operator over the caller's emotion as well as over the intensity and valence of emotion. First, we show the general architecture of the system. Then we

will discuss the different modules of the system's pipeline separately. The architecture relies on data acquisition and local as well as remote video data processing (figure 3).

The system runs some computer vision and AI operations on the mobile phone directly. Then, the facial expressions are analysed though cloud-based microservices.

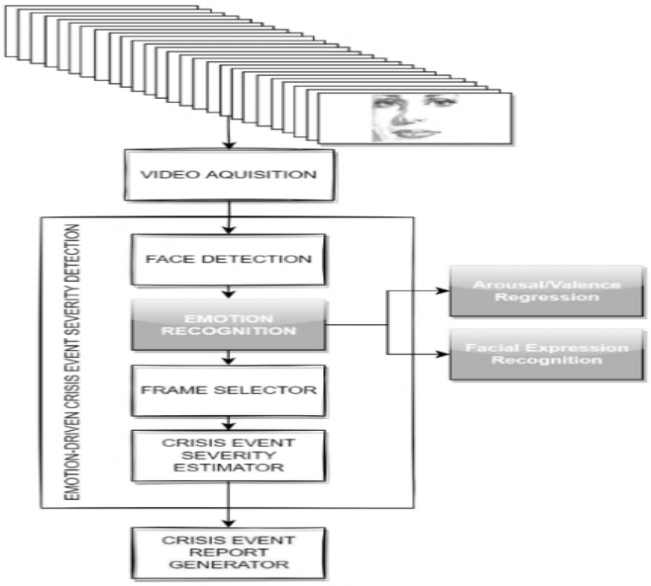

Fig. 3. Emotion-driven Emergency Incident Filtering Pipeline.

3.1. Video Acquisition

The **Video Acquisition** module (figure 3) connects the data processing pipeline with the device's video camera. A continuous sequence of video data is readily made available for live streaming and further processing.

3.2. Face Detection

The second module (**Face Detection**) runs on each video frame and detects the Region Of Interest (ROI) containing the caller's face. The module integrates the Viola & Jones algorithm [14] for detecting full view frontal upright faces in each input frame. A face is detected by extracting Haar features from grayscale images and by robustly spotting and linking patterns of face features such as nose, eyes, mouth. The underlying face detection uses Adaboost algorithm to simultaneously select relevant Haar features and learn how to link these visual features to the sub-patterns of a face. The result of applying the Viola&Jones face detector fully on an input video frame is represented by the list of rectangular ROIs of the detected faces.

3.3. Emotion Recognition

The next module is the **Emotion Recognition** module (see figure 3). In case the caller's face is visible and is detected by **Face Detection** module, the **Emotion Recognition** module gets the cropped sub-image showing the caller's face and reads the emotion. This module is actually running as a micro-service service in the cloud. That means the mobile application initiates requests to the cloud for processing each cropped face separately.

To automatize the process of reading emotion from faces, we built two emotion models. The first model estimates the emotion by assessing the arousal and the valence components of emotions. The second model integrates a classifier of the facial expressions.

Lately, Convolutional Neural Networks (CNNs) have been used extensively for facial expression recognition [10][11][12], providing more robust models when compared to the more traditional non-deep learning techniques.

We take a CNN pretrained on ImageNet, keep the first layers which account for the feature extraction, remove the classification and add a new regression module. The technique is called transfer learning.

Database Preparation

AffectNet [1] is by far the largest existing database of facial expressions, valence, and arousal in the wild enabling research in automated facial expression recognition in two different emotion models. For our research, we used a mini version of AffectNet that only contains the manually annotated images with 8 labels. That version of the database is labeled as AffectNet8 [1]. The image samples in AffectNet database are cropped and resized to 224 x 224 pixels (RGB color). The 8 expression labels and the arousal/valence values as well as the facial landmark points of the training and validation sets are in the database. The image samples are generated considering 15% boundary expansion of the face ROI, as provided through the OpenCV face detector. For our research, we used 8 emotion categorical labels as well as the valence and arousal values of the facial expressions in continuous domain. Because the part of AffectNet8 database used for this research shows highly unbalanced emotion categories, we resampled the database. The original AffectNet8 database contains 291.650 samples, out of which 287.651 are reserved for training and 3999 samples are for validation. The following steps were applied for resampling:

1) The samples for 8 emotion categories were rescaled to size 64x64 (for arousal-valence recognition) and separately to 128 x 128 pixels (for facial expression recognition), both for the train and validation datasets. The samples were kept the original color data channels.
2) The emotion categories under-represented were applied oversampling to generate new synthetic data. The emotion class contempt was augmented with 21k samples, disgust got 21k extra samples, fear was added 19k samples and surprise was extended with 11k extra samples. The data

augmentation applied a set of basic holistic image operations starting from the original samples. The operations implied rotating the whole image with ±5º (probability 50%), flipping the image horizontally (probability 50%), applying random brightness (probability 50%) and random contrast (probability 50%) or applying a full histogram normalization (probability 30%).

3) The emotion categories which were initially over-represented were applied under sampling to remove some of the samples. The emotion category happy got pruned 108k samples and emotion category neutral was reduced by 48k samples.
4) The train dataset and validation dataset were joined. We split the resulting dataset into three new datasets following the convention: 70% of the samples are for training, 15% for validation and 15% for testing.

The result consists of three separate datasets. The train dataset contained 145.126 samples. It consisted of 17.752 samples of category anger, 17.674 of contempt, 17.710 of disgust, 18.113 of fear, 18.718 of happy, 19.110 of neutral, 18.150 sad, 17.899 of surprise. Both the validation dataset and the test dataset had 31.093 samples each. Both data sets are structured as follows: 3.803 samples of anger, 3.786 of contempt, 3795 of disgust, 3880 of fear, 4011 of happy, 4094 of neutral, 3889 of sad and 3835 samples of surprise. The image samples in AffectNet8 were normalized before training and testing the affect models.

Arousal-Valence Regression Model

According to dimensional models, the emotion can be described using a dimensional space, most commonly the 2D space of valence and arousal. The valence of emotion represents the emotion as being positive or negative. The arousal dimension of emotion indicates the intensity or strength of the emotion state described [4]. To estimate the arousal and the valence, we built a CNN model for regression. The model was built by using first a VGG16 model and by adding our own regression layers on top of it. VGG16 is a convolutional neural network model (see figure 4) proposed by K. Simonyan and A. Zisserman from the University of Oxford in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition" [13]. The model achieves 92.7% top-5 test accuracy in ImageNet, which is a dataset of over 14 million images belonging to 1000 classes.

Next, we refined the model using the AffectNet8 emotion data. The top layers of VGG16 are trained to classify 1000 categories, whereas we need the model to classify 8 categories only. So, we removed the top module of VGG16 which accounts for classification. That is, we used the VGG16 module for feature extraction only. Next, we added a few more layers, as follows: a flattening layer, a dense layer (512 filters, ReLu activation) and a dropout layer (with 50% probability). The final layer is the regression layer with linear activation function. We train the newly added layers together with the top 5 layers of the original VGG16 model. The valence-arousal regression model has 15.764.802 parameters out of which 14.619.394 are trainable parameters.

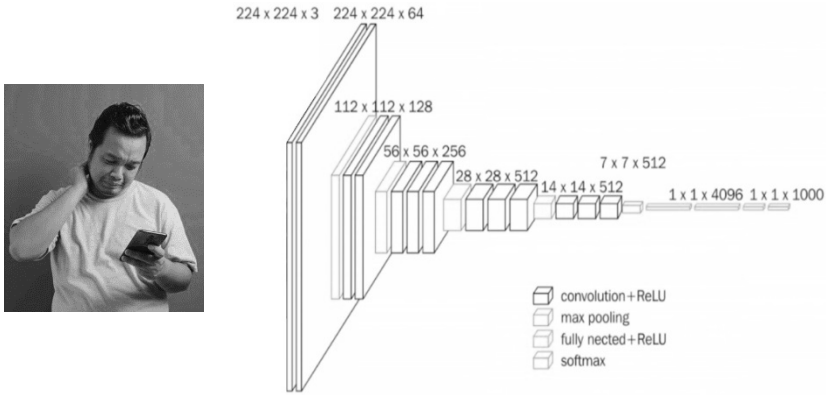

Fig. 4. The VGG16 convolutional neural network model for arousal-valence regression [1]

In AffectNet8, the values for arousal and valence were provided as floating-point numbers in the interval [-1,+1]. The regression model uses the Mean Squared Error loss function and SGD optimizer with learning rate 0.0001. Figure 5 depicts the loss curve obtained during training and validating the dual arousal-valence model.

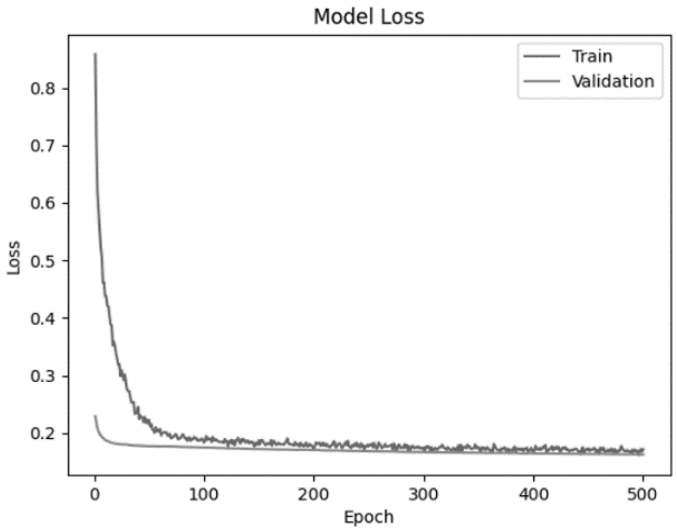

Fig. 5. The Arousal-Valence Regression Model Loss

Facial Expression Recognition Model

To recognize the facial expression of the emergency caller, we built a CNN classification model. The model was built by using VGG16 model, by removing the final layers and adding our own classification layers. The model was refined by training some existing layers and our new layers on AffectNet8 data. The model is

[1] Image from the left: Freepik.com

designed to learn and analyze neutral and seven categories of emotions. The new layers are as follows: a flattening layer, a dense layer (512 filters, ReLu activation) and a dropout layer (50% probability). The final layer is the regression layer with linear activation function. We train the newly added layers together with the top 5 layers of the original VGG16 model. The facial expression recognition model has 15.782.632 parameters out of which 8.146.280 parameters are trainable.

The left side of figure 6 shows the train and validation accuracy of the facial expression recognition model. The right side of figure 6 illustrates the loss curve of the same model, for train and validation.

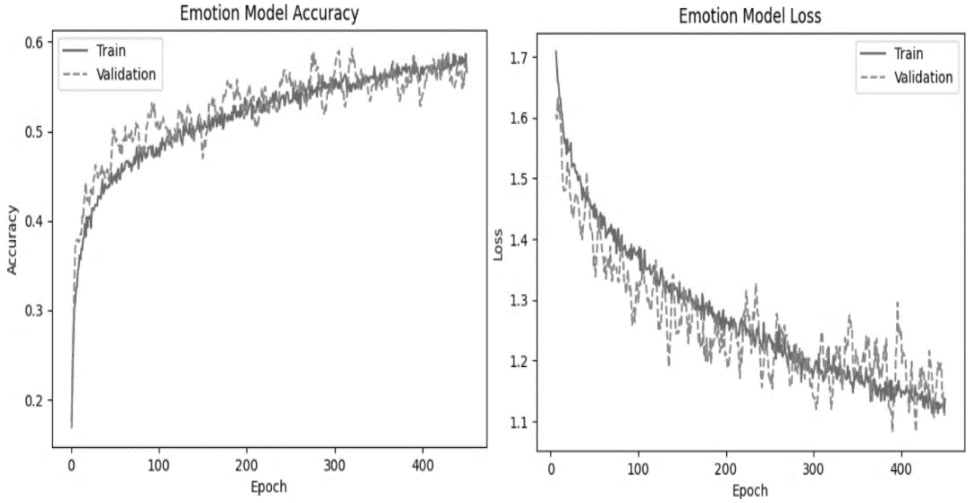

Fig. 6. The Accuracy (left) and Loss (right) for the Facial Expression Recognition Model

3.4. Frame Selector

Once the caller's face is detected and properly annotated using emotion labels, the module **Frame Selector** (see figure 3) filters the current set of face observations and choose the one which is the most relevant during the call. The filtering algorithm takes into account the result of both arousal and valence model and the categorical classifier of emotions. The most dominant facial expressions are further reported, their occurrence time, duration and the degree of mixing with other facial expressions of emotion. The system is trained to focus especially on the emotions with negative valence and also on the emotions expressed with higher intensity.

3.5. Crisis Event Severity Estimator

The selected frame segments together with the associated emotion readings are used to appreciate the severity of the emergency situation. This information is also passed further to the remote emergency operator, for better and quicker understanding the full emergency context.

4. DISCUSSIONS

The facial expression recognition model shows 60% accuracy on the test dataset. The accuracy rate of about 60% is somehow in line with the reduced performance of facial expression recognition when applied in real-world [15].

After building and testing the arousal-valence model and facial expression recognition model on AffectNet data, we have tested the models on real-life like emergency videos showing people expressing emotions while speaking on the phone. The results achieved were below the performance rates indicated during the testing stage, for both emotion models. That is mainly because of the influence of the face gestures triggered while speaking during the call. In addition, the light conditions from the real-life like situations differed considerably from those in the recordings of AffectNet database.

5. CONCLUSION

In this paper we described a system which reads the emotion of the caller to the emergency services and further assists the emergency personnel in better making decisions to handle the incident. The system integrates an application which runs directly on the smartphone of the caller and focuses on the caller's face from the live video data. The emotion reading is fully automatic and relies on assessing the valence and arousal of emotion, as well as reading categorical emotions. We made use of Transfer Learning and Convolutional Neural Networks (CNNs) to design the two emotion reading algorithms. Both models were trained using image samples from AffectNet database. The first results indicate good performance on reading the emotion from the face in emergency calls. The next step is to refine the emotion models by making them more robust to face gestures triggered while speaking.

ACKNOWLEDGMENT

We express our gratitude to Professor Mohammad Mahoor for providing access to AffectNet database.

REFERENCES

[1] Mollahosseini, A., B. Hasani, M.H. Mahoor. A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Transactions on Affective Computing*, 2017.

[2] Rothkrantz, L. Detecting Emotions in Communication via Social Media during Crisis Events. *2021 International Conference on Information Technologies (InfoTech)*, 2021, pp.1-8.

[3] Ekman, P., W.V. Friesen, P.Ellsworth. *Emotion in the human face: Guidelines for research and an integration of findings*. 1972, New York: Pergamon Press.

[4] Russell, J.A. Core affect and the psychological construction of emotion. *Theoretical Psychological Review*. Vol. 110, No. 1, 2003, pp.145–172.

[5] Wiegman, M. , J.Kersten, F.Khan, M.Pottharst, Stein. *Disaster Tweet Corpus*, 2020.

[6] Svensson, M., O. Pesämaa. How does a caller's anger, fear and sadness affect operators' decisions in emergency calls? *International Review of Social Psychology*, Vol. 31, No. 1, 2018, pp.7.

[7] Datcu, D., L. Rothkrantz. Facial expression recognition in still pictures and videos using active appearance models: a comparison approach. *CompSysTech '07: Proceedings of the 2007 International Conference on Computer Systems and Technologies*, June 2007, No.112, 2007, pp.1-6.

[8] Rothkrantz, L., S. Fitrianie. Public awareness and education for flooding disasters. *Crisis Management - Theory and Practice*, IntechOpen, 2018.

[9] Rothkrantz, L. Mitigation using emergent technologies. In: Teodorescu HN., Kirschenbaum A., Cojocaru S., Bruderlein C. (eds) *Improving Disaster Resilience and Mitigation - IT Means and Tools*. NATO Science for Peace and Security Series C: Environmental Security, 2014, pp 203-223.

[10] Choi, In-kyu, Ha-eun Ahn, Jisang Yoo. Facial expression classification using deep convolutional neural network. *Journal of Electrical Engineering and Technology*. Vol. 13, 2018, pp.485-492. 10.5370/JEET.2018.13.1.485, 2018.

[11] Liu, L., R. Jiang, J. Huo, J. Chen. Self-difference convolutional neural network for facial expression recognition. *Sensors* (Basel). Vol. 21, No. 6, Mar 2021, art. 2250.

[12] Akhand, M.A.H., S. Roy, N. Siddique, Md A.S. Kamal, T. Shimamura. Facial emotion recognition using transfer learning in the deep CNN. *Electronics,* vol. 10, 2021, art. 1036.

[13] Simonyan, K., A. Zisserman. *Very deep convolutional networks for large-scale image recognition*. 2014, arXiv:1409.1556.

[14] Viola, P., M. Jones. Robust real-time object detection. *International Journal of Computer Vision*, 2001.

[15] Ullah, A., J. Wang, M.S. Anwar, T.K. Whangbo, Y. Zhu. Empirical investigation of multimodal sensors in novel deep facial expression recognition in-the-wild, *Journal of Sensors*, 2021.

Information about the authors:

Dr.ir. Dragos Datcu has a PhD. in multi-modal emotion recognition defended in 2009 at TUDelft, The Netherlands. He works as an Innovator in Dutch IT industry.

Prof. Drs.Dr. Leon Rothkrantz is a (co-)author of more than 200 scientific papers on AI, speech recognition, multimodal communication and education.

Manuscript received on 29 October 2021