Fast and Reliable Emotions Detection Adapted for Driver Monitoring and Online Psychotherapy Sessions

Costin-Anton Boiangiu*, Marius-Eduard Cojocea, Robert-Costin Bercaru***, Mihai Bran****, Mihai-Lucian Voncila*, Nicolae Tarba*, Cornel Popescu*, George Culea*****

*Faculty of Automatic Control and Computers, University POLITEHNICA of Bucharest, 060042, Romania (Tel: 040-762-609-111; e-mail: costin.boiangiu@cs.pub.ro).
** Research and Development Department, OpenGov Ltd., Bucharest 011054, Romania (e-mail: marius.cojocea@opengov.ro)
*** Data Science and Machine Learning Department, SAP Romania Ltd., Bucharest 013714, Romania (e-mail: robert-costin.bercaru@sap.com)
**** Psychiatry, Coltea Hospital, Bucharest 030171, Co-founder AtlasHelp, Romania (e-mail: mihai@atlashelp.net)
*****Power Engineering and Computer Science Department, Faculty of Engineering, University VASILE ALECSANDRI of Bacau, 600115, Romania (e-mail: gculea@ub.ro)}

Abstract: In this paper, we present a solution for human face monitoring, which can be used in multiple scenarios. The presented solution monitors how a person felt throughout the whole therapy session, what was relaxing to talk about, what made him or her angry or disgusted. Thus, psychotherapists may acquire more data about their patients, in addition to what they already collected. Another use case is monitoring a car driver, based on their emotions and blinking patterns, to ensure that the driver is in a suitable state. This paper presents a method to assess the feelings a person has, in the domain of the five primary emotions: happiness, sadness, surprise, anger, and disgust. Besides emotions, our model is capable of monitoring how tired a person is, by monitoring their eyes and blinking patterns. To ensure that a high detection rate is performed, a machine learning approach based on a convolutional network was employed, backed up by a solid training phase performed onto a considerable set of tagged visual information. The proposed method compares favorably against other state-of-the-art emotion detection solutions for the proposed scenario and its performance is validated using a bespoke online psychotherapy image dataset acquired using low-end webcams with CMOS sensors. The results proved that the proposed solution is both fast and dependable, thus being currently used with strong results in a real-world platform for supporting psychologists in their remote counseling real-time sessions.

Keywords: emotion recognition; facial detection; automatic labeling; assisted driving; sleep detection; online psychotherapy; convolutional neural network; CNN.

1. INTRODUCTION

The use of patient monitoring during therapy can be motivated by multiple factors, one of them being the fact that some patients are not entirely open in their sessions with psychotherapists. They could avoid certain types of conversations or try to change the subject when it becomes unpleasant.

With the help of the presented solution and its subsequent results, one psychologist may review how the patient felt during certain moments of the therapy session and, eventually, reach a conclusion with a higher degree of confidence. In addition, the psychologist could adapt its techniques for the next sessions to provide better care for the patient.

Also, monitoring drivers can be a plausible solution for decreasing the number of car accidents. Thus, a car driver who is not in a proper condition for driving, due to being sleepy, distracted, or under the influence of some strong negative emotions, can be alerted.

The current paper's goal is tilted towards a practical, more applied solution, offering a helping hand for two real-world problems, online psychotherapy sessions and driver monitoring, and not just a theoretically inclined approach and discussion about emotion detection in general. Because the environments pertaining to these sorts of problems tend to be controlled, in general, the paper chooses not to deal with the issue of classifying emotions in the wild. There are a lot of other research and software solutions, many of them mentioned in this paper, which manage to answer the problem of emotion classification in a more general-purpose way. The short-term objective of this project is to create an artificial solution that can accurately tell how a person feels at a certain point in time and how the person's feelings vary during a period. The long-term objective is to find as many usage scenarios as possible.

The most obvious scenario is to create an intelligent assistant destined for helping psychologists during their remote counseling sessions, to automatically sample and log the status of their patient during one session or across multiple sessions

and to take the treatment decisions accordingly. A pilot phase of this artificial emotion detector solution is already running with strong results at AtlasHelp (AtlasHelp, 2020), an online counseling platform founded and conducted by the psychiatrist Mihai Bran. Some people may fear that lots of individuals are perfectly able to show a full range of external emotions without even changing their internal state. It's the case of actors, comedians, politicians, etc. As a result, any system developed specifically for the purpose may be tricked into delivering the wrong results. It may be true, it may even prove disastrous in the case of lie detector systems, but in the case of the psychotherapy sessions, either real-life or online, it is not a real issue. For a psychotherapy session to be successful, the subject must try as hard as possible to be fully open and sincere. In the end, it is in his best interest for the procedure to function according to the therapy plan.

Monitoring a driver's emotions (Lee et al., 2020) and level of rest (to identify if the driver is in a proper state for driving) is another obvious scenario for a fast and reliable emotion detector solution. Thus, a car computer could monitor the emotional state of the driver and offer assistance and indications when a driver is not in a suitable state for driving (e.g., detecting angry emotions). Also, by monitoring the driver's eyes, it could detect if the driver is keeping his eyes closed too much, which indicates that he or she is falling asleep. Thus, the driver can be alerted to this fact and may prevent serious accidents.

The presented solution can be later targeted to match other scenarios too, such as telling if people are lying based on their micro-expressions and helping innovative companies with actual feedback of what people are feeling during the launching event of a new product, extracting a global emotion from an image with a crowd of people, a personal monitor for home wellness and so on.

The first part of this paper presents some of the previous studies carried in this field. In the next part, the paper covers information about the technologies that were used during the implementation of the proposed solution after which information about the implementation itself is presented. The last part of the paper is dedicated to the results that were obtained using the proposed solution.

2. RELATED WORK

The idea that an artificial system is able to interpret genuine human emotions is not a new one. It gave birth to a lot of ingenious approaches, techniques, systems, and applications (Alonso-Martin et al., 2013; Rincon et al., 2019; Riaz et al., 2020). Apart from face image analysis (Turabzadeh et al., 2018), a lot of innovative approaches were proposed based on EEG BCI (Al-Nafian et al., 2017), miscellaneous biomarkers (Zamkah et al., 2020), speech (Sekkate et al., 2019), calibration games (Bevilacqua et al., 2019), as well as contact-based and skin-penetrating electrodes (Dzedzickis et al., 2020). iMotions and Emotient have worked together to create a software application for facial emotion recognition as well (Emotient, 2020; Mone and Sensing, 2015). The forementioned application not only does live-analysis with the help of a webcam but can also provide an emotion-related output based on previously recorded video(s) that can be uploaded to the iMotions servers. The analysis done by this software includes, but is not limited to, 7 basic emotions: joy, anger, surprise, fear, contempt, sadness, and disgust. Apart from these, it also offers data about 2 "advanced emotions" that are not considered "universal" according to Dr. Paul Ekman research group (Paul Ekman Group, 2020): confusion and frustration. Also, the application extracts an overall sentiment of a subject with positive, negative, and neutral streams. The background research is strongly based on FACS (Facial Action Coding System) (Imotions, 2020; Hamm et al., 2011; Ekman et al., 2002). The system refers to a set of different muscle movements and postures that directly correspond to an emotion. Although FACS dates to 2002, analyzing feelings was being done manually. Imotions does this automatically via their software.

The project developed by iMotions and Emotient is extremely popular and well-received in the community. It was even used on an episode of MythBusters (Benazzouz and Boudour, 2020) for analyzing the face of a driver while speaking on the phone. The results were critical in determining if a myth was true or not and were considered unquestionable. They combine this software with Stimuli, Facial Expressions, EEG, GSR, and more. Their final goal is to achieve the best accuracy possible in emotion detection. Even if techniques could be more invasive for the user than just looking at a camera, as seen in Figure 1, this is, as far as iMotions and Emotient are concerned, the way to go.

Fig. 1. The standard setup for emotion detection by iMotions and Emotient. Image retrieved from (Emotient, 2020).

Affectiva is another company with interests in this field, being focused more on market research and what they call "facial coding" (Affectiva, 2020; Magdin and Prikler, 2018). Their software solution measure 7 emotions: anger, contempt, disgust, fear, joy, sadness, and surprise. Apart from this, it also provides 20 facial expression metrics. Their algorithm uses Active Learning and an Efficient Non-Linear Kernel Approximation, as presented in (Senechal et al., 2015). The project started with the use of a "classic RBF SVM classifier", but the engineers noticed that this type of approach creates various problems in real-life scenarios, and it is not entirely reliable. Thus, a different approach was proposed with the hope of achieving greater performance. With this idea in mind, the team working on Affectiva paid close attention to the training data they used and switched to a Nystrom kernel approximation method. The goal was to have an algorithm that performs well during "spontaneous and naturalistic webcam videos". The specific action that Affectiva tracks for the moment consist of AU02 (outer eyebrow raise), AU04 (eyebrow lower), and smiles as seen in Figure 2. In the future,

the engineers plan to extend their software to track a greater number of facial action unit classifiers. They believe that their active learning approach will be even more effective when used on action units that appear less frequently than smiles and eyebrow movements. Also, another plan for the not-so-distant future of this team is to use the Stochastic Gradient Descent (SGD) optimization which could help in reducing the time needed for the algorithm to be trained and overcoming the memory limitations.

Fig. 2. Examples of AU02, AU04, and smiles. Image retrieved from (21Senechal et al., 2015).

Noldus's FaceReader (Noldus, 2020) is another advanced software solution for the detection of all the basic emotions. It is a complex and complete solution intended to offer not just emotion detection but also other characteristics such as head orientation, gaze direction, subject's gender, and age. It adapts its internal algorithms to the category of the subject: baby, child, adult, older person to provide better overall accuracy. It employs the Viola-Jones algorithm (Viola and Jones, 2004) for the face detection process, then it tries to reconstruct a 3D model of the face using a mesh determined by 500 key points and the application of the face's texture. Then it uses deep learning to perform face analysis and emotion classification. According to (Stöckli et al., 2018), FaceReader 6 was in certain scenarios, the best performer when compared against the major emotion classification tools available at that time.

One of the emotion detection solutions providers is NVISO (NVISO, 2020) which through its Insights line of products tries to detect and predict specific human behaviors using deep learning approaches. In terms of emotion classification, its work is based on the theoretical research of Dr. Paul Ekman (Ekman et al., 2002) enhanced using proprietary 3d facial imaging technology and advanced artificial intelligence methods.

CrowdSight F.A.C.E. API by Sightcorp (Sightcorp, 2020) is a complete solution for humans and crowd's behaviors and statistics, providing advanced Computer Vision and deep learning algorithms for face detection, crowd demographics, attention analysis, and, of course, emotion analysis. It also may provide various statistics like age and gender estimation, facial expressions, ethnicity, clothing style, head pose estimation, and general mood estimation.

Driver monitoring is increasingly affordable and plausible, due to the omnipresence of sensors in the modern world. It can be achieved by the car's sensors or using an independent device, such as a smartphone. (Castignani et al., 2015) present a solution that uses smartphone sensors to identify driving maneuvers, which can be used to create a driver profile to increase driving accuracy and reduce the risk of accidents and the number of risky maneuvers.

(Jo et al., 2011) propose a solution for monitoring driver drowsiness and distraction, based on the eye position and blinking.

Getting back to the presented research, in the beginning, a total of seven emotions were considered: anger, disgust, happiness, sadness, surprise, fear, representing the six basic emotions, and the neutral one. However, two emotions were discarded from the classification phase. During the experiments conducted with psychotherapists, it was concluded that the neutral emotion should not be considered for the detection process because it has too many similarities with all the other emotions and very few, if any, specific characteristics and, as a result, is not especially important in the psychotherapy decision process. Fear was also discarded due to the following reasons: it is an emotion that varies visually very much from person to person, or for the same person in various scenarios, it is much more difficult to be simulated by people, making the images labeled with this emotion not reliable for depicting the "actual fear" emotion, and – most importantly - it seldom appears in a natural psychotherapy session, especially when compared to the other emotions. Thus, the psychotherapists concluded that for an online system to be effective it must dependable detect only five emotions: happiness, sadness, surprise, anger, and disgust. The same considerations may apply when monitoring a driver's state.

The proposed solution employs the best architecture for the problem at hand and reduces the number of the detected emotions to just the five ones which are essential in a psychotherapy or driver monitoring process, to enhance the robustness of the classification when compared to other state of the art solutions.

3. PROPOSED SOLUTION

Since the data used for training consists of images, Convolutional Neural Networks (CNN) are an excellent choice for the model. The concept is around since at least 1998 when Yann LeCun et al. (LeCun et al., 1998) proposed a Convolutional Neural Network architecture for handwritten and machine-printed character recognition. Despite this, CNNs became popular and widely used since 2012, when such an architecture, called AlexNet (Krizhevsky et al., 2012) was the winner of the ImageNet Large Scale Visual Recognition Challenge 2012 (ImageNet, 2012) by a large margin. CNNs are usually the most recommended solution when working with images because they accept raw images as input, thus no spatial dependency between pixels is lost. Also, CNNs acts as automatic feature extractors, meaning that there is no need for extracting features in the preprocessing step. Many times, the features selected by humans are biased, which may lead to narrowing the learning range. CNNs learn their own features from the visual data used as input. These features may seem odd but can be used to perform image processing tasks with very good performance (Erhan et al., 2014; Ghaffari and Sharifian, 2016). Since the proposed solution is designed for computers that don't necessarily possess powerful GPUs, the architectures of the Convolutional Neural Networks used must be shallow, in order to keep the inference time low, ranging

from hundreds of milliseconds to a couple of seconds. The models developed for this solution have architectures like the CNN architecture of LeNet (LeCun et al., 1998; Ghaffari and Sharifian, 2016) and YOLO (Redmmon et al., 2016; Du, 2018). The initial model used for this project was inspired by the TensorFlow examples (Brownlee, 2017) and it was a simple, shallow architecture. The structure of this model can be seen in Figure 3.

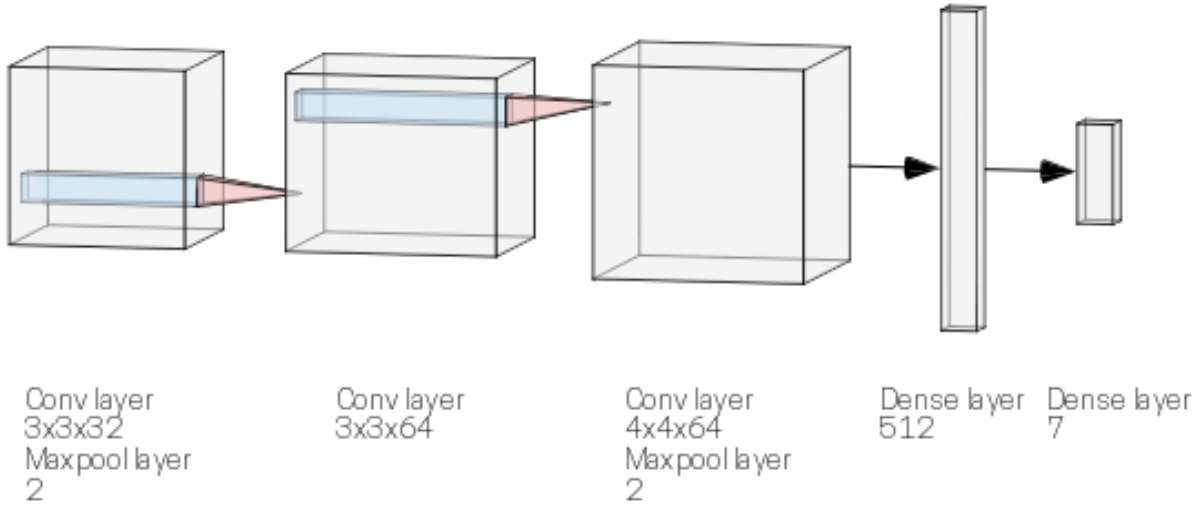

Fig. 3. The architecture for "Model 1".

The initial model suffered several changes regarding its layers and hyper-parameters, and it was trained for 50 epochs, with snapshots at every 10 epochs. The performance achieved during this phase was encouraging but could still be improved. This model will be further denoted as "Model 1" and references will be made to it later.

In the next phase, a deeper model was developed, which was inspired by YOLO (Brownlee, 2017). Its architecture can be seen in Figure 4.

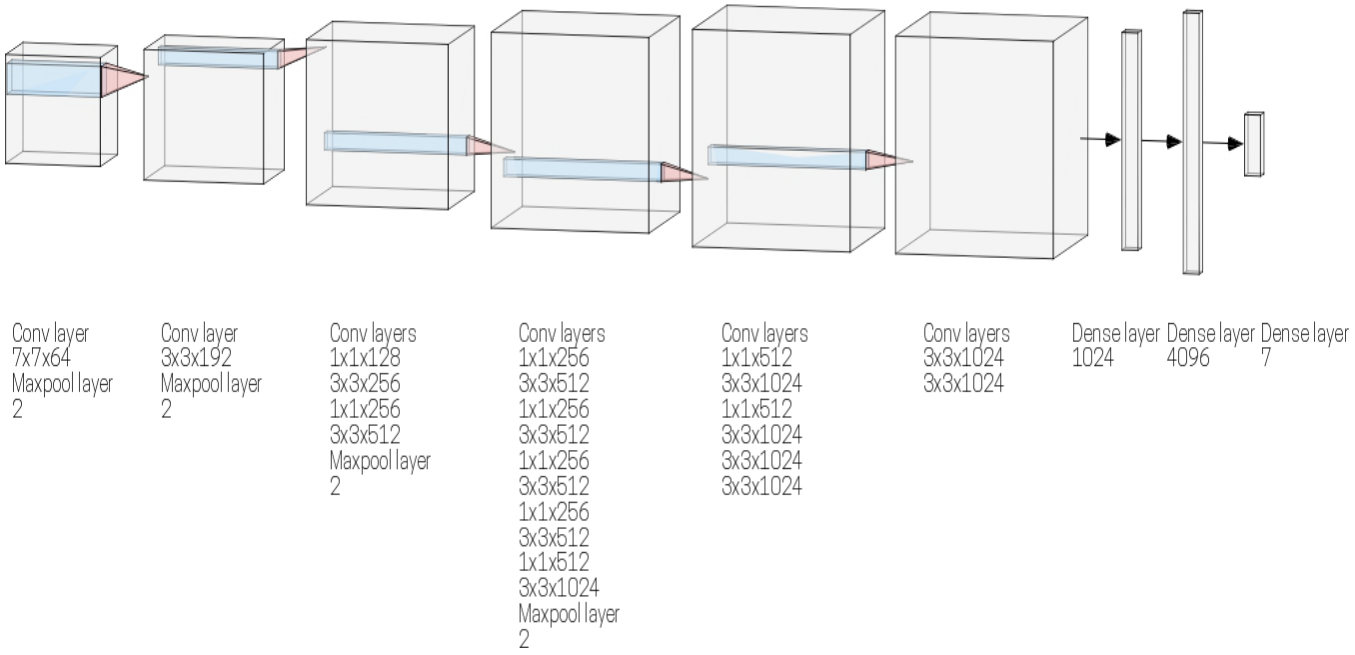

Fig. 4. The architecture for "Model 2".

The main difference between this model and the one described in YOLO is the absence of strides in the max-pooling layers. The model with the best results using this architecture will be further denoted as "Model 2". Despite using a deeper architecture, which can extract more features, the results were poorer. Thus, a new architecture was tried, shallower than Model 2, but deeper than Model 1, which is presented in Figure 5.

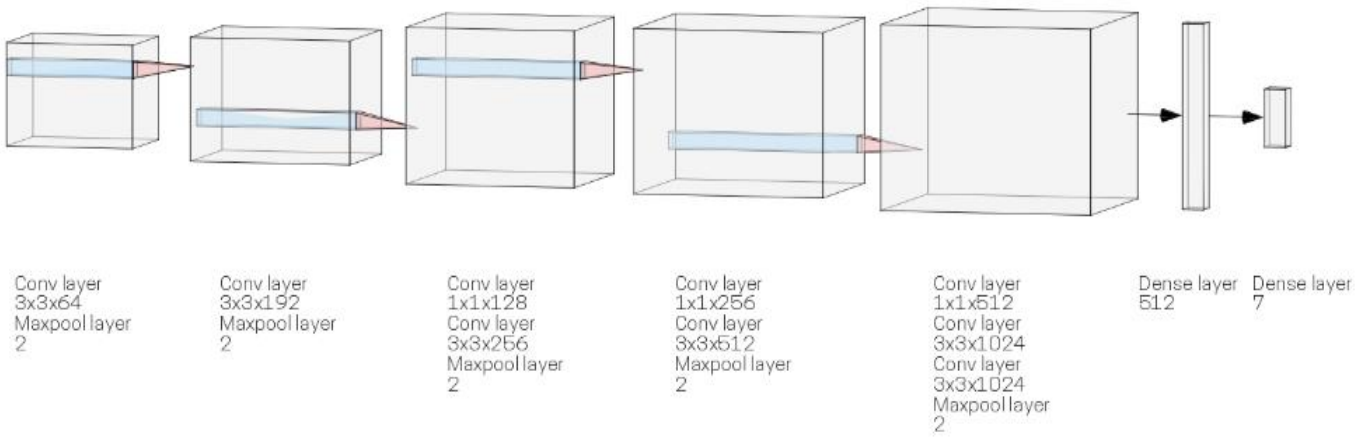

Fig. 5. The architecture for "Model 3".

This architecture was created by using the advantages of both the previous two networks. The structure is based on the second architecture, having more layers than the first, but the filters' sizes and numbers are like the ones used in the first architecture. Being reasonable in size, this network was trained for 50 epochs, and snapshots were saved every 10 epochs, just as with the first one. The best model obtained after the training will be further denoted as "Model 3".

The dataset which was used for these models is a mixture of the RAFD dataset (38Langner et al., 2010) and AffectNet dataset (Mollahosseini et al., 2017), containing more than 400 thousand labeled images, with people aged from 1 to 77 years, which was split in 80% for the training set and 20% for the validation set. It is to be noted that the AffectNet dataset contains two distinct kinds of annotations associated with each image, one representing a categorical model, which pertains to the basic emotions described by Ekman, and one representing a dimensional model, associated with valence and arousal. Of the two, only the first is of interest for this paper and is used in training the network. The AffectNet dataset also happens to contain images that might have poor lighting conditions, are occluded in some fashion, or present subjects in different postures, etc., as opposed to the RAFD dataset. Even though using such data could prove to be useful, it is not the main concern, since doctors can guide their patients' posture during online sessions, and give other specific instructions, whilst drivers are unlikely to change the way their camera faces, or their overall posture when driving.

A more specific test dataset (furtherly denoted as the "real-life" dataset) was created as well, by manually taking webcam snapshots of various people surprised in the same real-life conditions as in the scenario of an online psychotherapy session. The resulting pictures are labeled after the acquisition process in accordance with their corresponding emotions. The dataset consists of 200 pictures from 10 female and male subjects of various ages. All the persons in the real-life dataset are placed in relevant environments as during an online psychology session using a webcam, with respect to the lighting conditions and the subject posture and relative position in relation to his/her webcam. In the dataset creation process were employed 5 different low-end webcams with CMOS sensors, running in both native HD 720p resolution (1280x720) and reduced VGA (640x480).

We have implemented a CNN model for detecting the state of a person's eyes (e.g., open or closed). This is important since by using this model to sample the video stream at a frame rate of at least 10 fps, we could differentiate between regular blinking and falling asleep. This is since the duration of regular blinking averages between 100 and 400 msec. Thus, if too many successive detections of the eyes state are in the "closed" class, then the driver should be alerted to the fact that he or she is falling asleep.

For the blink detection model, we used the "Model 1" architecture, presented above, with minor changes, such as using a 3x3 filter in the last convolutional layer and a 64 dense layer after it. We have used the MRL Eye Dataset (MRL, 2020; Fusek, 2018), consisting of almost 85 thousand images of human eyes, labeled according to eye state (open, closed), gender, glasses, reflections, lighting conditions, and the quality of the data acquisition sensor.

4. IMPLEMENTATION DETAILS

The programming language used for developing the proposed solution was Python. Besides the ease of use, it was also chosen for its great support in image processing and machine learning libraries OpenCV (Howse, 2013), Tensorflow (Raschka and Mirjalili, 2019).

The training is done using labeled images, which will be employed by the neural network to learn to differentiate between the desired classes of emotions. The training of the model is time-consuming, ranging from hours to days, depending on the sizes of the dataset and the neural network. However, the inference time is small. Depending on the size of the neural network and the available hardware, the response time may vary between a few milliseconds to a few seconds. This is enough for the real-time detection of emotions during sessions with a psychologist.

All the models described in this paper require only images as input. No landmarks, or other elements, are used.

Since the proposed solution is aimed to be used on a normal computer, which does not have a high-end GPU, we had to consider only shallow networks, which contain only a few convolutional layers. This is necessary to keep inference time low. Deeper architectures, such as the deep variations from the YOLO family, which have tens or even above one hundred

layers, have very low inference time, measured in tens of seconds or even minutes, in the absence of a powerful GPU.

Thus, models having shallow architectures as presented in the previous chapter have been developed. When the detection process is started, the model is loaded from a file that contains the graph and the weights of the network. Afterward, the process is ready for inference.

The solution accepts input from a specific folder or from a video stream (webcam) (Bercaru, 2018). The output can be saved on the hard disk (used for manual labeling) or the solution can directly output an emotion classification. In both situations, the user must choose a sampling period at which the images are extracted. In the situation of having an online stream, the sampling will be done based on time, whereas if the input is from a specific folder (offline), the sampling will be done by establishing a specific frame rate. In the current state of the project, the sampling of the video stream from a webcam is being done every few seconds, by setting up a timer that calls the method responsible for the interrogation of the video stream, thus obtaining each frame. Each selected frame will be processed and sent over to the neural network to be analyzed for emotions.

As an alternative to sampling a video stream, a file-system watchdog was implemented. This will monitor a specific directory for new files and changes to existing files as well. When a new image or video file is created, it waits for it to be fully written on the disk. This is performed by inspecting the file size until two successive results are equal. If the file is an image, then the specific image is loaded, and it is being preprocessed to be fed to the neural network. Otherwise, if the file is a video, the watchdog extracts frames at each second of the video and creates images with them in the same input folder so they can be furtherly processed as a new image would be.

4.1 Preprocessing

The preprocessing step involves converting the image to grayscale, applying a face detection and extraction algorithm on each frame (Figure 6-a), followed by a resizing of the image to the resolution of 128x128 pixels if a face has been detected. (Figure 6-b).

(a) (b)

Fig. 6. Example of an image before the preprocessing step (a) and of the cropped, resampled, gray-scale photo obtained from the original (b).

Since all the processed faces are “squished” in the same manner and their necessary final resolution is a small one, no special precautions about the input image quality should be taken, since one’s facial area will always contain at least the needed number of pixels.

Also, the images are randomly flipped left right, rotated, and blurred, to achieve model invariance to affine transforms and to improve the generalization capability. The face extraction step is performed using trained Haar cascade classifiers (Paliy, 2008; Wilson and Fernandez, 2006). The image is first converted to grayscale, after which the algorithm marks the coordinates of a detected face (front or profile). The quality of the face extraction directly impacts the performance of the model, since it is desired that the input image given to the model contains as little of the background as possible. The images containing faces will be scaled to a given dimension by using a Lanczos4 interpolation (Madhukar et al., 2013).

4.2 Network response

After the analysis of the input data, the network will provide a dictionary, which consists of each possible emotion with its associated probability. These probabilities will sum up to 1.0. The emotion with the highest probability will be considered as the detected emotion.

For the alternative implementation that is using a watchdog, the results are being written to disk as a JSON file, as shown in Listing 1. The original file name will be the image id, and the JSON file will contain this id, which may be needed for future use, the timestamp at which the file was written, and the probability for each emotion.

Listing 1. An example of the network’s one-shot response as a JSON file.

```
{
   "time stamp": "2020-04-20 16:08:04",
   "emotions": {
      "anger", "0.0000",
      "sadness", "0.0000",
      "happiness”: "1.0000"
      "surprise”: "0.0000",
      "disgust”: "0.0000"
   }
   "id": "pixels-snapshot-713312"
}
```

Furthermore, the solution also provides the psychologist with graphs that shows the evolution of the emotions felt by a subject over certain periods of time.

The quick successions of states on short periods of time or shifts in subject’s states on longer periods may indicate the therapist that the working plan is adequate, and signs of progress are obvious or, perhaps, the changes are not the expected ones, and the therapy plan must be altered. Logging the graphs over longer periods of time may also help in objectively analyzing long-term improvements, performing diagnostics, or taking the appropriate decisions.

Figure 7 presents such a graph where, to emphasize visually the quick changes in the emotional state, the data is not obtained during a real-time online session but taken from a succession of photographs.

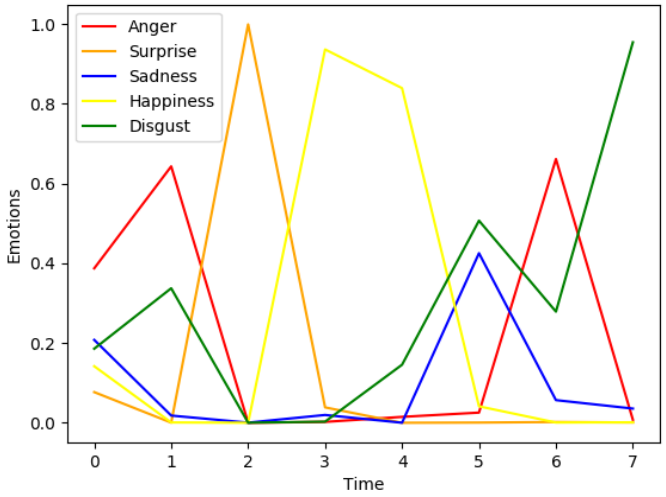

Fig. 7. An example of the network's response over time.

4. RESULTS

The results obtained are promising, showing that the task of robustly recognizing the essential emotions in a psychotherapy online session employing just low-end webcams is entirely feasible. All the employed webcams use cheap CMOS sensors, proving that better quality that comes with the more expensive and rarely used CCD sensors is not required.

The execution environment needs just a moderate computational power, due to the necessity that the proposed solution should be able to fully perform on a regular computer, without access to a GPU. This is necessary so that psychologists can use it without requiring any dedicated hardware investment. Thus, it is implied that the results can still be significantly improved if better hardware is available. The results obtained on the test and the real-life dataset are presented in Table 1.

Table 1. Precision, recall, and F1 score for the test and real-life dataset.

Emotions	Precision	Recall	F1 score
Anger	0.53	0.64	0.58
Disgust	0.51	0.63	0.56
Happiness	0.75	0.67	0.71
Sadness	0.49	0.43	0.46
Surprise	0.63	0.53	0.57
Average	0.58	0.58	0.58

Compared with nowadays high-end solutions for the same problem, the designed system performs very well for the online psychotherapy scenario reduced emotion set. It is important to note, however, that the one-on-one comparison is, somehow, advantageous to our solution since it is designed to track fewer emotions for improved detection accuracy and the solution was custom-tailored to perform ideally in environments that resemble online psychotherapy sessions developed in front of a webcam as it is the case with the employed dataset.

A comparison between the results of the three models described in Section 3 can be seen in Figure 6, where it is clear that "Model 3" is the best performing model.

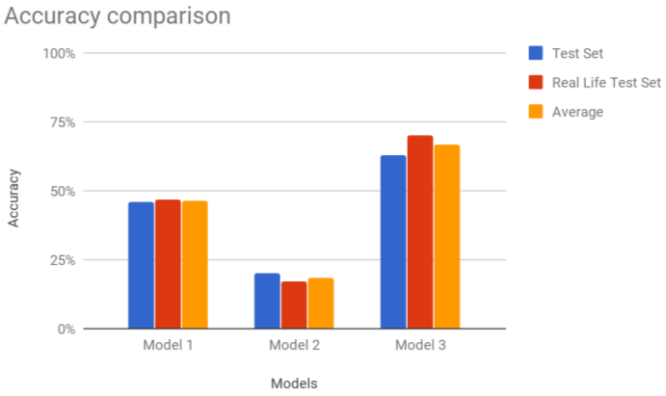

Fig. 8. The comparison between the accuracy of "Model 1", "Model 2" and "Model 3".

A comparison with the results of Microsoft Azure and Affectiva can be seen in Figure 9. Each of the above-mentioned systems was tested on the Real-Life test dataset mentioned earlier, which was designed specifically to mimic the input of real online psychotherapy sessions. Anyway, the comparison may not be performed on a larger test set taking into consideration the fact that, for each image, the demos of the state-of-the-art solutions had to be operated manually. Therefore, lacking the possibility of automatically testing a big batch of images, a smaller test set was used. The comparison could only be done with Microsoft and Affectiva's software solutions, because Imotions and NVISO do not have open demos, despite the requests made, and more recent non-commercial publications, such as those referenced, do not have available source code or binaries to test the dataset.

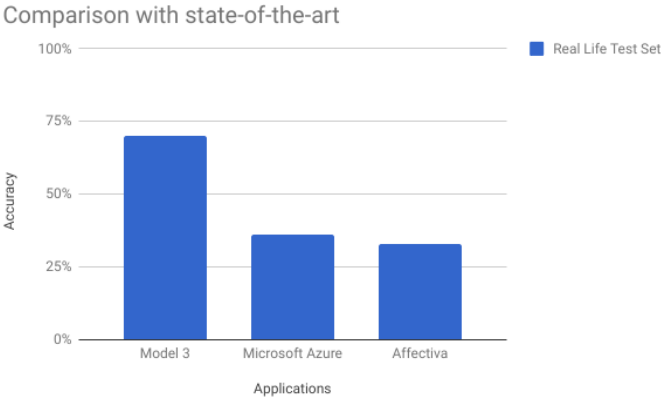

Fig. 9. The comparison in detection accuracy, where the presented model achieved around 70% accuracy.

Regarding the accuracy that the three systems provided, it is worth saying that the two state-of-the-art applications track more than the 5 basic emotions detected by this project and sometimes get confused between the "neutral" state and the "real" emotion expressed. Another thing that was quite surprising while testing these systems was that the demos from Microsoft and Affectiva have poor performance on face detection, which could have a significant impact on the emotion classification performance. Even though the project described in this paper possesses a higher accuracy, it is worth noting that Microsoft's solution is a much more complex one than just tracking emotions. It is part of a much bigger project, as was mentioned in section 3 and it also tracks three more

emotions, which can result in a higher error rate. Also, Microsoft designed its system as a collection of APIs that can be used by developers all over the world. In terms of speed, Affectiva's system is much faster than the one from Microsoft. Using the Affectiva's application on a mobile device, it instantly reads emotions as one's face enters the frame of the camera (and, of course, if the face is successfully detected), compared to Microsoft's and the proposed solution, which may take at least half a second to produce any results.

Table 2 displays the results obtained for the eye state detection model. One can see that it is a high-performance model, using 100x100 pixels images of eyes as input and running at more than 30 frames per second on a machine with Intel® Core™ i7-7820HK CPU @ 2.90 GHz, 2904 MHz, 4 Core(s), and 8 Logical Processor(s). Thus, it is a model that can be reliably used in practical scenarios for detecting when the driver of a car is falling asleep.

Table 2. Accuracy, precision, recall, and F1-score for eye state detection.

Accuracy	Precision	Recall	F1
98.54%	98.10%	99.04%	98.57%

The obtained results proved strong, thus achieving the goal of the current research, which was to create a solution for emotion detection that can be a useful tool for psychologists and a valuable helper for safe driving. The presented solution may be significantly improved both in terms of speed and performance when better hardware is available, both for training, but especially for inference. Also, in the psychotherapy scenario, the results may be improved by enriching the dataset with more images and by online training (using feedback from psychologists).

Regarding the ethical issues about detecting emotions, we all agreed that no such technology should be used without the consent of the monitored person. We do not imply or recommend that it should be used otherwise. The main focus was set on using it for driver monitoring and therapy sessions, and only with the explicit consent of the subject, as is the case now in the proposed solution.

6. CONCLUSIONS

The research presented in this project may result in a significant improvement in various fields of work. The short-term objective was to create a fast and reliable solution that could detect five of the basic emotions in a manner that could be used by professional psychotherapists in their sessions with patients. This would come in the aid of the specialists to determine the actual emotions that one patient experiences. The resulting counseling platform is in no way intended to become a replacement for the psychology sessions, but rather a tool that can improve the psychologists' work with the patients. Also, it is a tool that could save many lives by preventing accidents caused by people driving when they are tired or in a bad emotional state. Thus, the car's response to such a situation could vary from simple alerts and notifications to blocking the car until the driver is able to drive it safely.

6.1 Future developments

One of the possible improvements is to deal better with the face occluding cases, such as people wearing glasses. An idea that can be very useful is to create a new neural network that automatically removes the glasses from a face contained in an image and run all the input images to this network first, and then through the one that the project already has. A proof of concept that this is feasible can be found in (Wu et al., 2004).

Other architecture improvements could imply deeper networks, different architectures, network pipelines, parallel processing pipelines. Furthermore, increasing the size of the training data could result in a significant increase in performance. Moreover, online training could be useful, where the persons using this artificial emotion detector solution will notify when the model makes classification errors on real-life data (Simpson, 2015).

Another possible direction may be towards data exchanging between vehicles and their manufacturers (Campanile et al., 2020) to gather significant information about drivers' habits and the evolution of their alertness during long journeys, thus allowing the manufacturers to implement better safety procedures.

REFERENCES

Affectiva Emotion AI. Available online: https://www.affectiva.com/emotion-ai-overview/ (accessed on July 10, 2020).

Al-Nafjan, A.; Hosny, M.; Al-Ohali, Y.; Al-Wabil, A. Review and Classification of Emotion Recognition Based on EEG Brain-Computer Interface System Research: A Systematic Review. Appl. Sci. 2017, 7, 1239.

Alonso-Martín, F.; Malfaz, M.; Sequeira, J.; Gorostiza, J.F.; Salichs, M.A. A Multimodal Emotion Detection System during Human-Robot Interaction. Sensors, 2013, Vol. 13, 15549-15581.

AtlasHelp - Counselling – Anytime & Anywhere. Online or offline meetings with certified and licensed Specialists, Available online: https://atlashelp.net/ (accessed on July 10, 2020).

Benazzouz, Y.; Boudour, R. An Emotion-Based Search Engine. In: Arai K., Bhatia R., Kapoor S. (eds); Proceedings of the Future Technologies Conference (FTC) 2019. Advances in Intelligent Systems and Computing, 2020, vol. 1069, Springer, Cham, DOI: 10.1007/978-3-030-32520-6_15.

Bercaru, R.C. Aplicație pentru măsurarea emoțiilor primare. License Thesis, Unpublished work, Bucharest, 2018.

Bevilacqua, F.; Engström, H.; Backlund, P. Game-Calibrated and User-Tailored Remote Detection of Stress and Boredom in Games. Sensors 2019, 19, 2877.

Brownlee, J. Deep Learning with Python: Develop Deep Learning Models on Theano and TensorFlow Using Keras. Machine Learning Mastery, Melbourne, Australia, 2017.

Campanile, L.; Iacono, M..; Maruli, F.; Mastroianni, M. Privacy Regulations Challenges on Data-centric and IoT Systems: A Case Study for Smart Vehicles. Proceedings of IoTBDS 2020, the 5th International Conference on

Internet of Things, Big Data and Security, 2020, Vol. 1, pp. 507-518, DOI: 10.5220/0009839305070518.

Castignani, G.; Derrmann, T.; Frank, R.; Engel, T. Driver behavior profiling using smartphones: A low-cost platform for driver monitoring. IEEE Intelligent Transportation Systems Magazine, 2015, Vol. 7, No. 1, pp. 91-102, DOI: 10.1109/MITS.2014.2328673.

Du, J. Understanding of Object Detection Based on CNN Family and YOLO. Journal of Physics: Conference Series. IOP Publishing, 2018, pp. 012029.

Dzedzickis, A.; Kaklauskas, A.; Bucinskas, V. Human Emotion Recognition: Review of Sensors and Methods. Sensors 2020, 20, 592.

Ekman, P.; Friesen, W.V.; Hager, J.C. Facial Action Coding System - The Manual on CD-ROM. Salt Lake City: A Human Face Publisher, 2nd edition, 2002.

Emotient and iMotions Partner to Offer Unique Integrated Facial Expression Recognition, Bio Sensor and Eye Tracking Solution for Usability, Gaming, Market and Academic/ Scientific Research. Available online: https://www.prnewswire.com/news-releases/emotient-and-imotions-partner-to-offer-unique-integrated-facial-expression-recognition-bio-sensor-and-eye-tracking-solution-for-usability-gaming-market-and-academic-scientific-research-210226401.html (accessed on July 10, 2020).

Erhan, D.; Szegedy, C.; Toshev, A.; Anguelov, D. Scalable object detection using deep neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR'14), 2014, pp. 2147-2154.

Fusek, R. Pupil localization using geodesic distance. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Vol. 11241, 2018, pp. 433-444, DOI: 10.1007/978-3-030-03801-4_38.

Ghaffari, S.; Sharifian, S. FPGA-based convolutional neural network accelerator design using high level synthesize. 2nd International Conference of Signal Processing and Intelligent Systems (ICSPIS), Tehran, 2016, pp. 1-6. DOI: 10.1109/ICSPIS.2016.7869873.

Hamm, J.; Kohler, C.G.; Gur R.C.; Verma, R. Automated Facial Action Coding System for dynamic analysis of facial expressions in neuropsychiatric disorders. Journal of Neuroscience Methods, September 2011, Vol. 200, No. 2, pp. 237-256, DOI: 10.1016/j.jneumeth.2011.06.023.

Howse, J. OpenCV Computer Vision with Python. Packt Publishing Ltd, 2013.

Imotions – Facial Action Coding System presented by Imotions. Available online: https://imotions.com/blog/facial-action-coding-system/ (accessed on July 10, 2020).

Jo, J.; Lee, S.J.; Kim, J.; Jung, H.G.; Park, K.R. Vision-based method for detecting driver drowsiness and distraction in driver monitoring system. Optical Engineering, 2011, Vol. 40, No. 12, 127202, DOI: 10.1117/1.3657506.

Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 2012, pp. 1097-1105.

Kuo, J. Understanding convolutional neural networks with a mathematical model. Journal of Visual Communication and Image Representation, November 2016, Vol. 41, pp. 406-413. DOI: 10.1016/j.jvcir.2016.11.003.

Langner, O.; Dotsch, R.; Bijlstra, G.; Wigboldus, D.H.J.; Hawk, S.T.; van Knippenberg, A. Presentation and validation of the Radboud Faces Database. Cognition & Emotion, 2010, Vol. 24, No. 8, pp. 1377–1388, DOI: 10.1080/02699930903485076.

ImageNet – Large Scale Visual Recognition Challenge 2012 (ILSVRC2012) Available online: http://www.image-net.org/challenges/LSVRC/2012/ (accessed on July 10, 2020)

Lee, S.; Lee, T.; Yang, T.; Yoon, C.; Kim, S.-P. Detection of Drivers' Anxiety Invoked by Driving Situations Using Multimodal Biosignals. Processes 2020, 8, 155.

LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, Vol. 86, No. 11, pp. 2278-2324.

Madhukar, B.N.; Narendra, R. Lanczos Resampling for the Digital Processing of Remotely Sensed Images. Proceedings of International Conference on VLSI, Communication, Advanced Devices, Signals & Systems and Networking (VCASAN), 2013, pp. 403-411. DOI: 10.1007/978-81-322-1524-0_48.

Magdin, M.; Prikler, F. Real time facial expression recognition using webcam and SDK Affectiva. International Journal of Interactive Multimedia and Artificial Intelligence (IJIMAI), 2018, Vol. 5, No. 1, pp 7-15.

Mollahosseini, A.; Hasani, B.; Mahoor, M.H. AffectNet: A New Database for Facial Expression, Valence, and Arousal Computation in the Wild. IEEE Transactions on Affective Computing, August 2017, DOI: 10.1109/TAFFC.2017.2740923.

Mone, G. Sensing emotions. Communications of the ACM, 2015, Vol. 58, No. 9, pp. 15-16, DOI: 10.1145/2800498.

MRL Eye Dataset. Available online: http://mrl.cs.vsb.cz/eyedataset (accessed on July 10, 2020).

Noldus – Facial expression recognition software. Available online: https://www.noldus.com/facereader (accessed on July 10, 2020).

NVISO. Insights Advise and Insights Develop by NVISO – Fintech Artificial Intelligence. Available online: https://www.nviso.ai/en (accessed on July 10, 2020).

Paul Ekman Group - Universal Emotions. Available online: https://www.paulekman.com/universal-emotions/ (accessed on July 10, 2020).

Paliy, I. Face detection using Haar-like features cascade and convolutional neural network. International Conference on Modern Problems of Radio Engineering, Telecommunications and Computer Science (TCSET), Lviv-Slavsko, 2008, pp. 375-377.

Raschka, S.; Mirjalili, V. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2. Packt Publishing Ltd, 2019.

Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. IEEE

Conference on Computer Vision and Pattern Recognition (CVPR'16), Las Vegas, NV, 2016, pp. 779-788. DOI: 10.1109/CVPR.2016.91.

Riaz, M.N.; Shen, Y.; Sohail, M.; Guo, M. eXnet: An Efficient Approach for Emotion Recognition in the Wild. Sensors 2020, 20, 1087.

Rincon, J.A.; Costa, A.; Carrascosa, C.; Novais, P.; Julian, V. EMERALD—Exercise Monitoring Emotional Assistant. Sensors 2019, Vol. 19, 1953.

Sekkate, S.; Khalil, M.; Adib, A.; Ben Jebara, S. An Investigation of a Feature-Level Fusion for Noisy Speech Emotion Recognition. Computers 2019, 8, 91.

Senechal, T.; McDuff, D.; el Kaliouby, R. Facial Action Unit Detection Using Active Learning and an Efficient Non-linear Kernel Approximation, IEEE International Conference on Computer Vision Workshop (ICCVW), Santiago, 2015, pp. 10-18. DOI: 10.1109/ICCVW.2015.11.

Sightcorp, CrowdSight F.A.C.E. API by Sightcorp - Leverage the power of Face Analytics in your app! Available online: https://face-api.sightcorp.com/ (accessed on July 10, 2020).

Simpson, A.J.R. On-the-Fly Learning in a Perpetual Learning Machine. arXiv preprint, 2015, arXiv: 1509.00913, 2015.

Stöckli, S.; Schulte-Mecklenbeck, M.; Borer, S.; Samson, A.C. Facial expression analysis with AFFDEX and FACET: A validation study. Behavior Research Methods, Vol. 50, No. 4, August 2018, pp.1446-1460, DOI: 10.3758/s13428-017-0996-1.

Turabzadeh, S.; Meng, H.; Swash, R.M.; Pleva, M.; Juhar, J. Facial Expression Emotion Detection for Real-Time Embedded Systems. Technologies 2018, 6, 17.

Viola, P.; Jones, M.J. Robust real-time face detection. International Journal of ComputerVision (IJCV), Vol. 57, 2004, pp. 137–154, DOI: 10.1023/B:VISI.0000013087.49260.fb.

Zamkah, A.; Hui, T.; Andrews, S.; Dey, N.; Shi, F.; Sherratt, R.S. Identification of Suitable Biomarkers for Stress and Emotion Detection for Future Personal Affective Wearable Sensors. Biosensors 2020, 10, 40.

Wilson, P.I.; Fernandez, J. Facial feature detection using Haar classifiers. Journal of Computing Sciences in Colleges, 2006, Vol. 21, No. 4, pp. 127-133.

Wu, C.; Liu, C.; Shum, H.Y.; Xu, Y.Q.; Zang, Z. Automatic Eyeglasses Removal from Face Images. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2004, Vol. 26, No. 3, pp. 322-336, DOI: 10.1109/TPAMI.2004.1262319.