Avrupa Bilim ve Teknoloji Dergisi
Özel Sayı 32, S. 690-695, Aralık 2021

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
Special Issue 32, pp. 690-695, December 2021

Research Article

EduFERA: A Real-Time Student Facial Emotion Recognition Approach

Kaouther Mouheb[1], Ali Yürekli[2], Nedzma Dervisbegovic[3], Ridwan Ali Mohammed[4], Burcu Yılmazel[5*]

[1] Eskişehir Technical University, Faculty of Engineering, Department of Computer Engineering, Eskişehir, Turkey, (ORCID: 0000-0002-8991-9405), kmouheb@eskisehir.edu.tr

[2] Eskişehir Technical University, Faculty of Engineering, Department of Computer Engineering, Eskişehir, Turkey, (ORCID: 0000-0001-8690-7559), aliyurekli@eskisehir.edu.tr

[3] Eskişehir Technical University, Faculty of Engineering, Department of Computer Engineering, Eskişehir, Turkey, (ORCID: 0000-0003-3739-5336), nedzmadervisbegovic@eskisehir.edu.tr

[4] Eskişehir Technical University, Faculty of Engineering, Department of Computer Engineering, Eskişehir, Turkey, (ORCID: 0000-0002-5029-2887), ridwanalimohammed@eskisehir.edu.tr

[5*] Eskişehir Technical University, Faculty of Engineering, Department of Computer Engineering, Eskişehir, Turkey, (ORCID: 0000-0001-8917-6499), byurekli@eskisehir.edu.tr

(International Conference on Design, Research and Development (RDCONF) 2021 – 15-18 December 2021)

(**DOI:** 10.31590/ejosat.1039184)

ATIF/REFERENCE: Mouheb, K., Yürekli, A., Dervisbegovic, N., Mohammed, R. A., & Yılmazel, B. (2021). EduFERA: A Real-Time Student Facial Emotion Recognition Approach. *European Journal of Science and Technology*, (32), 690-695.

Abstract

The use of video conferencing tools in education has increased dramatically in recent years. Especially after the COVID-19 outbreak, many classes have been moved to online platforms due to social distancing precautions. While this trend eliminates physical dependencies in education and provides a continuous educational environment, it also creates some problems in the long term. Primarily, many instructors and students have reported issues concerning the lack of emotional interaction between participants. During in-place education, the speaker receives immediate emotional feedback through the expressions of the audience. However, it is not possible to fully utilize this valuable feedback in online lectures since current tools can only display a limited number of faces on the screen at a time. In order to alleviate this problem and promote the online education experience one step closer to in-place education, this study presents EduFERA that provides a real-time emotional assessment of students based on their facial expressions during video conferencing. Empirically, several state-of-the-art techniques have been employed for face recognition and facial emotion assessment. The resulting optimal model has been deployed as a Flask Web API with a user-friendly ReactJS frontend, which can be integrated as an extension to current online lecturing systems.

Keywords: Computer vision, Affective computing, Facial emotion recognition, Video conferencing, Online education.

EduFERA: Gerçek Zamanlı Öğrenci İfadelerinden Duygu Tanımlama Yaklaşımı

Öz

Son yıllarda video konferans araçlarının eğitim alanında kullanımında oldukça önemli bir artış gerçekleşmiştir. Özellikle COVID-19 salgını döneminde uygulanan sosyal mesafe tedbirleri, birçok dersin çevrimiçi platformlarda yürütülmesini gerektirmiştir. Bu trend, fiziksel bağımlılıkları ortadan kaldırarak sürdürülebilir bir eğitim ortamı sağlarken, uzun vadede bazı problemleri de beraberinde getirmiştir. Birçok eğitimci ve öğrencinin belirttiği üzere, çevrimiçi derslerde katılımcılar arası duygusal etkileşimde bir eksiklik yaşanmaktadır. Yüz yüze eğitim süreçlerinde, bir konuşmacı hitap ettiği kitleden anlık duygusal bir geri bildirim alabilmektedir. Öte yandan, video konferans araçları belirli bir zaman diliminde ekranda kısıtlı sayıda yüz gösterebilmekte; bu durum da dersin işleyişi açısından oldukça önemli olan duygusal geri bildirimden tam anlamda faydalanılamamasına neden olmaktadır. Bu çalışmada duygu etkileşimi probleminin üstesinden gelmek ve çevrimiçi eğitim deneyimini yüz yüze eğitim kalitesine yaklaştırmak amacıyla geliştirilen EduFERA tanıtılmaktadır. EduFERA modeli, video konferansı esnasında öğrencilerin yüz ifadelerini gerçek zamanlı işleyerek duygusal değerlendirmelerde bulunmayı hedefler. Modelin geliştirme sürecinde literatürdeki etkin yüz tanıma ve yüz ifadesinden duygu çıkarma yaklaşımları incelenmiştir. Deneysel sonuçlar ve analizler sonucunda elde edilen optimum model, Flask Web API aracılığında kullanıcı dostu bir ReactJS arayüzü ile mevcut çevrimiçi eğitim sistemlerine bir eklenti olarak hizmete açılmıştır.

Anahtar Kelimeler: Bilgisayarlı görü, Duygusal hesaplama, Yüz ifadesinden duygu tanıma, Video konferans, Çevrimiçi eğitim.

[*] Corresponding Author: byurekli@eskisehir.edu.tr

1. Introduction

In the era of rapid and dynamic business life, video conferencing tools play an important role in connecting people instantly. With the emergence of the COVID-19 outbreak, these tools have been also widely used in the field of education (Xie et al., 2020). Globally, many classes have been moved to online platforms due to social distancing measures aimed at preventing the spread of the pandemic. This way, it became possible to continue education in a crisis environment whose duration and scope were unpredictable. On the other hand, the shift towards online learning has also brought some particular problems. Many concerns about the lack of emotional interaction have been reported by both instructors and students (Baber, 2020; Aguilera-Hermida, 2020). Accordingly, the instructors cannot provide sufficient emotional interaction with the audience due to some natural software limitations, such as the limited number of faces displayable on screens and visual interruption during screen sharing. The lack of emotional interaction resulting from these limitations creates concerns in terms of the efficiency of online lecturing.

One possible way to alleviate the above-mentioned problem is to decorate video conferencing tools with the ability of facial emotion recognition (FER), which is the task of recognizing users' emotional states through their facial expressions (Picard, 2000). Empowering these tools with real-time FER on student video footage might help the instructors better engage with the audience and have more control over the lecture flow.

In this study, we present EduFERA, which is a real-time student emotion assessment approach that aims to improve the online lecturing experience. Inspired by the effectiveness of deep learning in correlated computer vision tasks (Farrell et al., 2019; Zhou et al., 2020; Zeng et al., 2020), the proposed system utilizes deep neural network architectures for face detection and FER. Several experiments are carried out on a common, real-world dataset. All these experiments are conducted on Google Colab[†] platform using GPU acceleration.

The main contributions of the study can be summarized as follows:

- An emotional assessment approach is proposed as a real-time feedback mechanism for students' engagement in online lectures.
- Several deep neural network architectures for face detection and FER are evaluated in an experimental setting.
- The resulting optimal FER model is deployed in a Web application for demonstration purposes.

The rest of the paper is organized as follows. In Section 2, the utilized materials and the methodology of EduFERA are introduced. In Section 3, the experimental results for face detection and FER are presented. Finally, conclusions are drawn and some future works are discussed in Section 4.

2. Material and Method

[†] https://colab.research.google.com/

2.1. The Circumplex Model of Affect

The majority of FER approaches investigate human emotions under seven primary emotional states (Zhou et al., 2020), which are anger, disgust, fear, sadness, happiness, surprise, and neutral. In the scope of this study, the main focus is the assessment of students' engagement and impressions about the content covered during an online lecture. Therefore, we utilize the circumplex model of affect (Russell, 2003) when transitioning from emotional states to the level of interaction between participants.

According to the circumplex model of affect, any particular emotion can be expressed as some factors of valence (i.e., pleasant to unpleasant) and arousal (i.e., activation to deactivation). Figure 1 illustrates the emotion space in terms of valence and arousal quadrants.

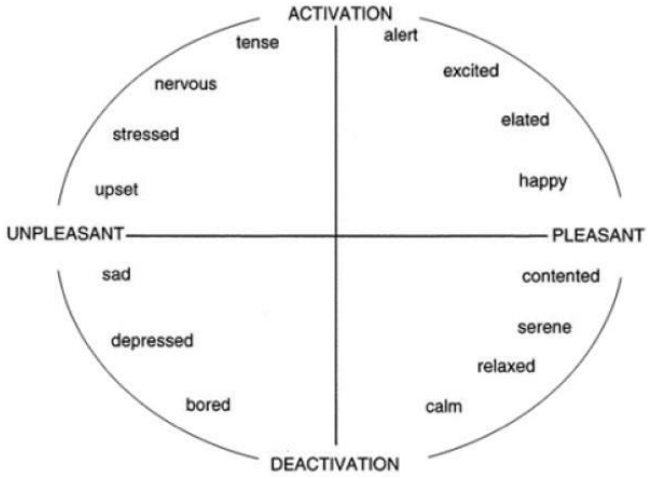

Figure 1. A graphical representation of the circumplex model of affect (Russell, 2003)

The quadrants visualized in Figure 1 can be described and exemplified as follows:

- **Active-Pleasant (AP)**: Emotional states with positive valence and positive arousal (e.g., curiosity and interest).
- **Active-Unpleasant (AU)**: Emotional states with negative valence and positive arousal (e.g., confusion and anger).
- **Inactive-Unpleasant (IU)**: Emotional states with negative valence and negative arousal (e.g., boredom and tiredness).
- **Inactive-Pleasant (IP)**: Emotional states with positive valence and negative arousal (e.g., satisfaction and ease).

When classifying student emotions, a two-dimensional vector having valence and arousal values as its features can be used to represent each emotion. Alternatively, the emotion space can be divided into four classes, where each class represents one quadrant.

2.2. The FER2013+ Dataset

In this study, the FER2013+ dataset (Barsoum et al., 2016) is utilized to train FER models. The FER2013+ is a re-labeled version of the FER2013 dataset (Goodfellow et al., 2013) using crowdsourcing methodology during image annotation. In total, the dataset consists of 35,886 images.

Each of the images in the FER2013+ is classified by ten different labelers into eight major emotion classes, which are “Happy”, “Surprise”, “Anger”, “Disgust”, “Fear”, “Sad”, “Contempt”, and “Neutral”. The images for which a consensus cannot be reached are tagged as “Unknown”. Similarly, the images without any faces are classified as “No Face”.

During the preprocessing phase of the FER2013+, we first remove the images that cannot be matched with a certain emotion (i.e., the images having “Unknown” or “No Face” labels). Then, we map the emotions to the quadrants described in the circumplex model of affect. The mapping is performed by using the association given in Table 1.

Table 1. Emotions in the FER2013+ dataset and their corresponding core affect

Emotion in FER2013+	Corresponding Core Affect			
	AP	AU	IU	IP
Happy	✓			
Surprise	✓			
Anger		✓		
Disgust		✓		
Fear		✓		
Sad			✓	
Contempt			✓	
Neutral				✓

The mapping between emotions in the FER2013+ dataset and core affects results in a highly imbalanced data collection. Therefore, we follow a balancing strategy that involves up-sampling on minority classes and down-sampling on majority classes. Consequently, each class is adjusted to contain 5000 training instances, 500 validation instances, and 441 test instances. In Figure 2, the resulting distribution after resampling is presented.

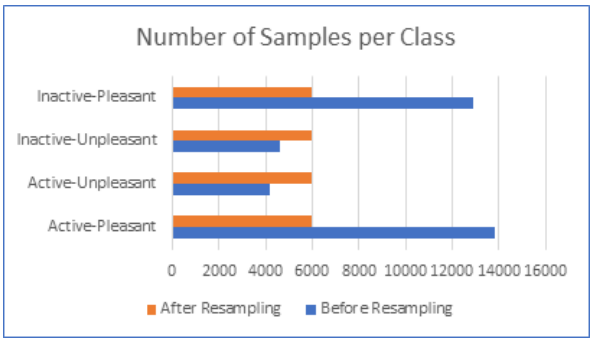

Figure 2. Number of samples per class before and after resampling

2.3. EduFERA Architecture

In order to alleviate the lack of emotional interaction problem in online lecturing, we propose EduFERA, which is a FER system for video conferencing tools. The proposed system consists of four main stages:

i. Face detection
ii. Image transformation
iii. Emotional classification
iv. Data visualization

The main purpose of the EduFERA is to increase the quality of the online learning experience by providing real-time emotional feedback to instructors. The overall architecture of the proposed system is presented in Figure 3.

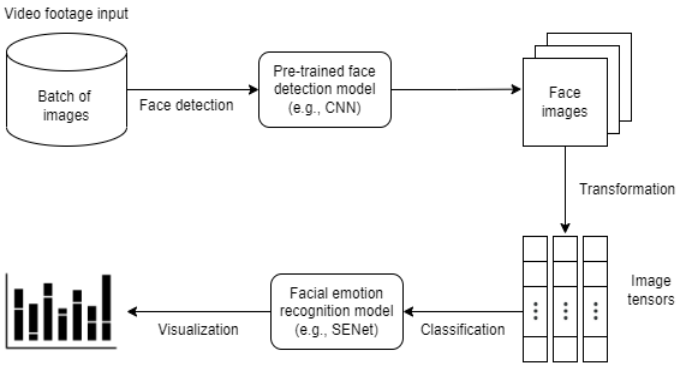

Figure 3. Architectural overview of EduFERA as an emotional assessment model

As illustrated in Figure 3, the system is designed to receive video footage as the primary input. Initially, the input is processed as a batch of images. Then, this batch is fed to a pre-trained face detection model. For each image, the facial section is cropped and passed to a transformation layer. In this layer, the images are grayscaled, resized to 224*224 dimensions, and normalized to the interval of [-1, 1]. After the transformation, the resulting image tensors are given to a trained FER model for emotional classification. The identified emotions are unified with respect to the circumplex model of affect. This way, a student’s engagement in the lecture can be estimated. Finally, the results are interpreted and presented to the target user of the system (e.g., an instructor or a lecturer).

For face detection and emotion recognition, we employ several state-of-the-art deep learning approaches (Goodfellow et al., 2013; Zeng et al., 2020) and focus on the optimal solution in terms of classification accuracy and processing power. In line with this goal, extensive experiments are carried out with a well-known facial images dataset (Simonyan & Zisserman, 2014). The experimental results, best-performing learning models in our setting, and our final product are introduced in the following sections.

3. Results and Discussion

3.1. Face Detection

The natural first step of FER is to recognize faces in an input image. The performance of face detection has a direct impact on the overall FER process. To equip EduFERA with a fast processing face detector, we evaluate four pre-trained models in terms of their processing times for the entire dataset. Briefly, the detectors rely on the following approaches:

- Convolutional neural networks (CNN)
- Multi-task cascaded convolutional networks (MTCNN)
- Max-margin CNN (MMOD CNN)
- HOG + Linear SVM

Table 2. Average processing times for the four different face detection models

Face Detection Model	Processing time per frame (GPU)	Processing time per frame (CPU)
MTCNN	**0.0144**	**0.0605**
HOG + Linear SVM	-	0.1983
MMOD CNN	0.0925	-
CNN	0.0865	0.105

Table 2 presents the average processing time of each face detector per image. According to the table, MTCNN performs much better than the other models in both GPU and CPU settings. While MTCNN processes a frame in an average of 0.06 seconds with the CPU, it speeds up approximately 6 times with the GPU, reducing the processing time to 0.01 seconds. Therefore, we employ MTCNN as the final face detection model of EduFERA.

3.2. Emotional Assessment

Given an input image, EduFERA identifies faces using a pre-trained face detection model. After a series of normalization steps (e.g., grayscaling and resizing), the resulting image tensors are fed to a deep neural networks architecture for emotional assessment. For an effective FER process in EduFERA, we evaluate the following four popular approaches in terms of accuracy and processing power:

- InceptionResNet V1 (Schroff et al., 2015)
- VGG-VD-16 (Albanie & Vedaldi, 2016)
- ResNet-50 (Albanie et al., 2018)
- SENet (Albanie et al., 2018)

Table 3. Comparison of four different FER models in terms of accuracy

FER Model	Validation accuracy	Test accuracy	F-score
InceptionResNet V1	0.7778	0.7582	0.76
VGG-VD-16	0.8213	0.7844	0.79
ResNet-50	0.8393	0.8125	0.81
SENet	**0.8562**	**0.8292**	**0.83**

Table 4. Comparison of four different FER models in terms of processing power

FER Model	Frames per second (GPU)	Frames per second (CPU)
InceptionResNet V1	207.2617	5.9545
VGG-VD-16	**260.4785**	**8.7362**
ResNet-50	193.9312	4.8957
SENet	180.4991	4.4538

Table 3 and Table 4 present comparisons of the four FER models in terms of accuracy and processing power, respectively. In the accuracy aspect, SENet achieves greater success than InceptionResNet V1, VGG-VD-16, and ResNet-50. However, in a real-time environment, image processing performance is also a critical factor. VGG-VD-16 outperforms the other models with its processing capacity of approximately 260 frames per second with the GPU. The difference is so significant that the slightly low accuracy of VGG-VD-16 is negligible. Therefore, we prefer VGG-VD-16 as the optimal FER model of EduFERA due to real-time performance concerns.

The architecture of VGG-VD-16 originates from Simonyan & Zisserman (2014). In a succeeding study, Albanie & Vedaldi (2016) train the network using stochastic gradient descent (SGD) (Sutskever et al., 2013) as the optimizer with a mini-batch size of 64 and a learning rate of 0.0001. Figure 4 illustrates the chart of the model's validation and training accuracy per epoch.

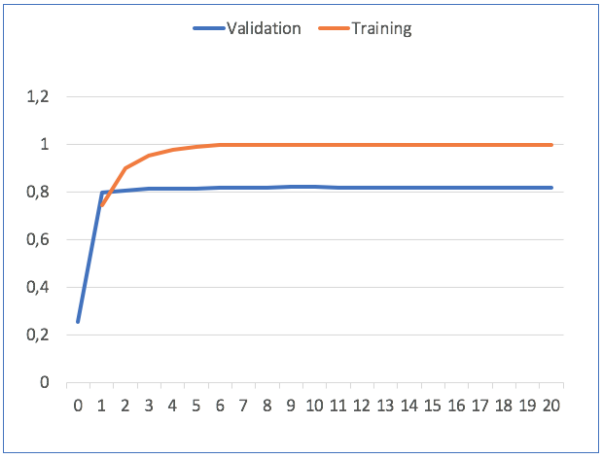

Figure 4. Validation and training accuracy of VGG-VD-16 per epoch

As the final evaluation of EduFERA, we analyze our combined FER model in terms of classification accuracy on the test set. As previously mentioned in Section 2.2., our resampling strategy in FER2013+ results in a total of 441 test instances. The confusion matrix given in Table 5 summarizes the emotional assessment in terms of the circumplex model of affect.

Table 5. Confusion matrix for the classification task of emotions with respect to the circumplex model of affect

	AP	AU	IU	IP
AP	381	27	10	23
AU	25	358	29	29
IU	13	37	305	86
IP	15	17	66	343

3.3. EduFERA as an End Product

In order to productize EduFERA and serve as an extension to existing video conferencing tools, we develop a Web application that is capable of receiving user requests, assessing emotional engagements based on the trained face detection and FER models (i.e., MTCNN and VGG-VD-16), visualizing the predictions, and storing results for later use. An architectural overview of the application is presented in Figure 5.

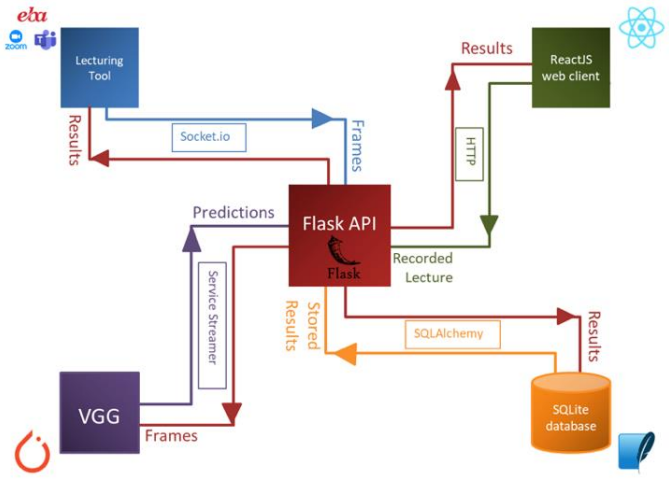

Figure 5. Architectural overview of EduFERA as an end product

In backend development, Flask API[‡] is used to establish data transfer between machine learning models and the application. For data visualization, a user-friendly interface is developed using ReactJS[§]. The primary features of the Web application can be summarized as current meeting analysis, recorded meeting analysis, individual tracking, and past meetings history. An example case of current meeting analysis is presented in Figure 6.

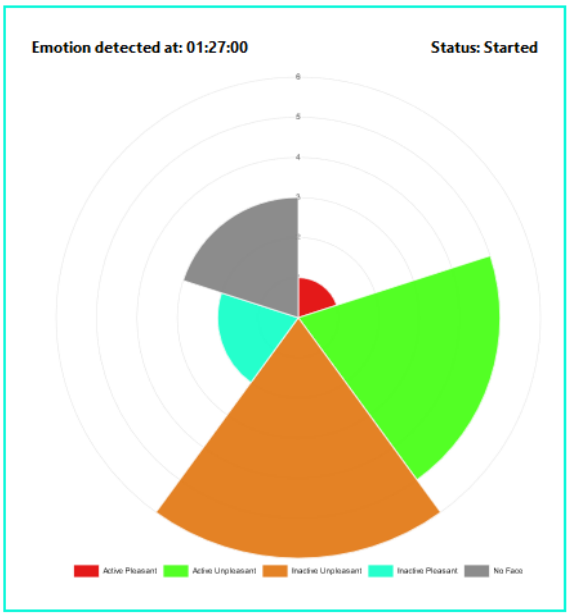

Figure 6. A sample visualization for emotional assessment from the EduFERA Web application

4. Conclusions and Recommendations

The lack of emotional interaction between the speaker and the audience is a critical problem in online education. Although video conferencing tools eliminate physical barriers in education, both instructors and students need elegant software assistance for a better educational experience.

This study presents EduFERA, which is a real-time emotion assessment extension for video conferencing tools used in online education. When learning face and emotion embeddings, several deep neural network architectures are evaluated and analyzed. The experimental results show that while MTCNN stands out among the other approaches in terms of processing power in face detection, VGG-VD-16 appears to be the optimal model for FER in terms of accuracy in a reasonable time. Employing MTCNN and VGG-VD-16 as the face detector and emotion recognizer, respectively, a Web application for EduFERA is also developed. In the near future, we will focus on performance enhancements for FER through transfer learning strategies. Furthermore, we are also planning to deploy EduFERA as an active Web application for public use.

5. Acknowledge

This study was supported by TUBITAK 2209-A under the grant no: 1919B012001659 and Eskişehir Technical University Scientific Research Projects Commission under the grant no: 21LTP030.

References

Aguilera-Hermida, A. P. (2020). College students' use and acceptance of emergency online learning due to COVID-19. International Journal of Educational Research Open, 1, 100011.

Albanie, S., & Vedaldi, A. (2016). Learning grimaces by watching tv. arXiv preprint arXiv:1610.02255.

Albanie, S., Nagrani, A., Vedaldi, A., & Zisserman, A. (2018). Emotion recognition in speech using cross-modal transfer in the wild. Proceedings of the 26th ACM International Conference on Multimedia, (s. 292-301).

Baber, H. (2020). Determinants of students' perceived learning outcome and satisfaction in online learning during the pandemic of COVID-19. Journal of Education and e-Learning Research, 7(3), 285-292.

Barsoum, E., Zhang, C., Ferrer, C. C., & Zhang, Z. (2016). Training deep networks for facial expression recognition with crowd-sourced label distribution. Proceedings of the 18th ACM International Conference on Multimodel Interaction, (s. 279-283).

Farrell, C. C., Markham, C., & Deegan, C. (2019). Real-time detection and analysis of facial features to measure student engagement with learning objects. IMVIP 2019: Irish Machine Vision & Image Processing.

Goodfellow, I. J., Erhan, D., Carrier, P. L., Courville, A., Mirza, M., Hamner, B., . . . Bengio, Y. (2013). Challenges in representation learning: a report on three machine learning contests. International Conference on Neural Information Processing, (s. 117-124).

Picard, R. W. (2000). Affective computing. MIT Press.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. Psychological Review, 110(1), 145-172.

Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: a unified embedding for face recognition and clustering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (s. 815-823).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional neural networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[‡] https://flask.palletsprojects.com/en/2.0.x/

[§] https://reactjs.org/

Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the importance of initialization and momentum in deep learning. International Conference on Machine Learning (s. 1139-1147). PMLR.

Xie, X., Siau, K., & Nah, F. F.-H. (2020). COVID-19 pandemic - online education in the new normal and the next normal. Journal of Information Technology Case and Application Research, 22(3), 175-187.

Zeng, H., Shu, X., Wang, Y., Wang, Y., Zhang, L., Pong, T.-C., & Qu, H. (2020). EmotionCues: emotion-oriented visual summarization of classroom videos. IEEE Transactions on Visualization and Computer Graphics, 27(7), 3168-3181.

Zhou, W., Cheng, J., Lei, X., Benes, B., & Adamo, N. (2020). Deep-learning-based emotion recognition from real-time videos. International Conference on Human-Computer Interaction (s. 321-332). Cham: Springer.