A New Facial Expression Processing System for an Affectively Aware Robot

Engin Baglayici[1](✉), Cemal Gurpinar[1], Pinar Uluer[1,2], and Hatice Kose[1]

[1] Istanbul Technical University, Istanbul, Turkey
{baglayici17,gurpinarcemal,hatice.kose}@itu.edu.tr
[2] Galatasaray University, Istanbul, Turkey
puluer@gsu.edu.tr

Abstract. This paper introduces an emotion recognition system for an affectively aware hospital robot for children, and a data labeling and processing tool called LabelFace for facial expression recognition (FER) to be employed within the presented system. The tool provides an interface for automatic/manual labeling and visual information processing for emotion and facial action unit (AU) recognition with the assistant models based on deep learning. The tool is developed primarily to support the affective intelligence of a socially assistive robot for supporting the healthcare of children with hearing impairments. In the proposed approach, multi-label AU detection models are used for this purpose. To the best of our knowledge, the proposed children AU detector model is the first model which targets 5- to 9- year old children. The model is trained with well-known posed-datasets and tested with a real-world non-posed dataset collected from hearing-impaired children. Our tool LabelFace is compared to a widely-used facial expression tool in terms of data processing and data labeling capabilities for benchmarking, and performs better with its AU detector models for children on both posed-data and non-posed data testing.

Keywords: Facial expression recognition · Action unit recognition · Child-robot interaction · Affective computing

1 Introduction

The interpretation of human behavior is highly dependent on emotions. The ability to recognize emotions is essential in psychology, education, human-robot interaction, healthcare, entertainment, and other fields that are related to human behavior. Facial expressions are the central cues to emotional inferences [18] and facial expression recognition (FER) is one of the most used methods for emotion recognition. FER is the process of extracting the features of facial expressions that show the clues of different emotions.

A. Del Bimbo et al. (Eds.): ICPR 2020 Workshops, LNCS 12662, pp. 36–51, 2021.
https://doi.org/10.1007/978-3-030-68790-8_4

Ekman and Friesen worked on facial expressions and the emotions, and defined six basic universal emotions such as, happiness, sadness, surprise, disgust, fear, and anger [6]. Based on the studies carried out on adults and children, it was observed that particular facial expressions correspond to particular emotions across all the cultures. The Facial Action Coding System is developed to distinguish emotions from facial expressions by categorizing the motions of facial muscles into action units (AU) [8]. Action units consist of contractions and relaxations of one or more muscles. They can provide an interpretation of facial expression by single atomic AU or a group of AUs (For AUs and their elemental and compound emotion categories, please refer to [12]).

The facial emotion recognition process involves data gathering, processing, labeling, and finally, training and testing. Data labeling can be an exhaustive process due to the amount of data that is required for the training of deep learning-based models [9]. Organizing and labeling this amount of data requires the automation of the process. Although researchers have used several data annotation tools for similar applications, most data annotation tools, such as ELAN and LabelMe, let the user manually supervise the labeling process. Furthermore, these tools are incapable of labeling several frames in video data in an autonomous manner. Lastly, these tools are general-purpose programs that do not provide a field-specific user interface.

We are working on the RoboRehab project consisting of an emotion recognition tool and a robotic system, specially developed for hearing-impaired children to be used during their audiometry tests in hospitals [28]. The main purpose of this study is to develop a system to recognize the emotions of children who are in interaction with a social and assistive humanoid robot. Unfortunately recognizing children's emotions especially in the wild is a very challenging task. We are in the search for a robust and feasible solution and decided to use AUs for emotion recognition. Also for young children, who have difficulty in showing their emotions, detection of AUs might be a helpful approach for affective systems.

During this study, we had to process, and label an extensive amount of video recordings from the robot and cameras for emotion recognition models, hence the need for a ready-to-use tool that covers automatic, effortless data labeling and processing. This paper introduces our proposed tool LabelFace, and we provide the pipeline and implementation details of such a tool, that researchers can use for their specific areas of interest.

LabelFace aims to contribute to the community with the services which, open-source tools do not fully provide, which are: 1) Enabling manual or automatic data labeling; 2) Manual or automatic image and video processing; 3) Ready to use emotion assistant models 4) Real-time performance 5) API for programmers 6) Open-source code.

For multi-label AU detection, we trained different models using the transfer learning method. We fine-tuned a generalized facial expression model, VGG-FACE [26], for specialized multi-label action unit detection task. To the best of our knowledge, children AU detector model is the first model that targets 5- to 9- year old children, especially for hearing-impaired children.

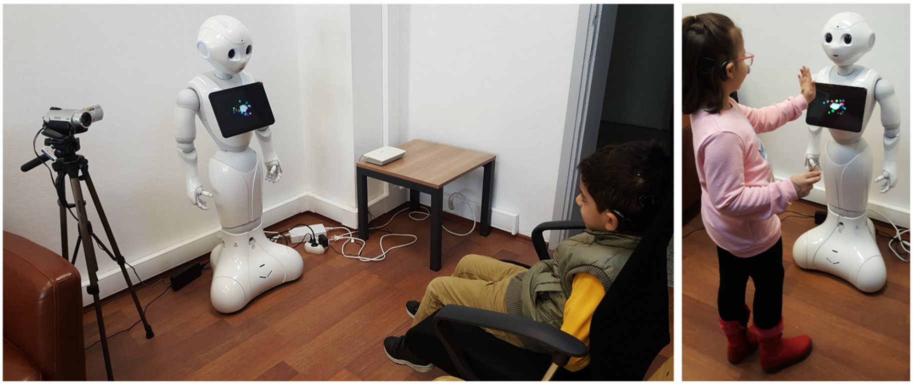


Fig. 1. CRI study with children

For benchmarking, we also examined the capabilities of another mostly used open-source tool, OpenFace 2.0. We compared action unit detection performances of both tools in terms of non-posed and posed datasets of children. We produced the non-posed data from a previous child-robot interaction (CRI) study with children having hearing aid or cochlear implant (Fig. 1).

For comparison, we included nine common AUs which are described in the following sections in details.

2 Related Work

Recognizing emotions is possible through examining information such as facial expressions, speech, text, biological (e.g., EEG) data [14]. Researchers tried to make use of this information either with single-model or multi-modal approaches [27]. Even though Ekman and other researchers claimed that some basic emotions are universal and common, others have argued that people generally do not react to the situations with the same emotional level and expressions [20].

Generally, conventional Facial Emotion Recognition (FER) workflow includes three main steps: visual information processing, feature extraction from the processed visual information, and classification of extracted features. In the visual information processing phase, various image processing operations are used to remove geometric variations between frames and utilize frames in a single format. Feature extraction can be done in different ways that are geometric-based feature extraction, appearance-based feature extraction, or a combination of these methods [16]. Geometric-based features are extracted based on the position and angle of facial landmark points. Appearance-based feature extraction makes use of essential face regions and the patterns in these regions. Finally, classification takes place with learning algorithms such as Support Vector Machine (SVM), Adaboost, Random-Forest.

In the past years, conventional computer vision methods are applied for FER, and successful results are achieved [1]. Recently, deep learning-based emotion-recognition systems have shown state-of-the-art performance, and several setups of these models are presented and implemented [4,13,23,27].

Deep learning eliminates the feature extraction phase from the FER problem by providing end-to-end learning processes and reduces mathematical complexity. Deep learning methods require large amount of data, and significant amount of computing power to process these data.

Convolutional Neural Network is the most used deep learning methods for detection and recognition tasks. Researchers utilized different CNN architectures for FER and got promising results after conventional approaches. In the CNN approach, images are convolved through a filter collection to extract a feature map. These feature maps are then combined to fully connected layers to classify incoming images [17].

Breuer and Kimmel used the CNN architecture to examine and demonstrate the essential features for FER and the relations of these features with the Facial Action Coding System (FACS) and Action Units [4]. Mollahosseini et al. propose a CNN model to address the FER problem and evaluates the model on multiple well-know standard face datasets [23]. The utilized CNN model is supported by inception layers to improve local feature performance. The model requires lesser computational requirements and provides increased accuracy in both subject-independent and cross-database evaluations.

CNN approaches are suitable for extracting spatial features, but they lack finding temporal variations in data [11]. Thus, other than CNN, Recurrent Neural Networks (RNN) and Long-Short-Term-Memory (LSTM) Networks [31] are also utilized for the FER problem.

LSTM and RNN architectures are specifically designed to find temporal features from time-series data. LSTM networks showed best performances with the video sequences [12]. Jain et al. present a hybrid CNN-RNN architecture that handles both spatial and temporal dependencies in images [13]. First convolutional layers are utilized to extract spatial features from frames. The extracted features are then passed to the RNN layers, which are connected to CNN layers serially. Extracting temporal features in RNN layers provides a significantly increased accuracy.

Action unit (AU) detection methods are generally divided into two categories, static-based and dynamic based methods [19]. Static-based methods interest in the spatial information of data to find patterns while dynamic-based methods interest in the temporal relations in data. Promising results are achieved by combining static and dynamic based methods in recent years [7]. A more detailed survey on deep learning with FER can be seen in [19].

FER studies on different age groups bring different challenges. For the children, it is not possible to include all action units, since facial features in children are not fully developed. Hammal et al. [9] proposed a multi-label CNN model to detect action units in infants. They made use of Baby FACS [24], which is an extension of FACS for infants. They emphasize the requirement of automatic detection of AUs in infants to address the needs of researchers and clinicians in this area [9].

There are various open-licensed or commercial facial expression recognition tools to analyze facial data from videos or still images (for a detailed survey on the open-licensed software, please refer to [3]). Commercial tools offer the user a large variety of machine learning-based solutions, such as basic emotion detection (Kairos[1], SkyBiometry[2], Findface[3], FaceReader[4]). FaceReader claims to be the first tool that can recognize facial expressions in infants.

OpenFace 2.0 [3] is a comprehensive open-licensed tool accompanied by facial landmark/action unit detection, head pose, and eye-gaze estimation features. OpenFace detects the presence of action units and their intensity values. OpenFace uses Support Vector Machines (SVM) to recognize AUs, and it uses only adult datasets [3]. For each AU, they trained a separate SVM model. Also the preprocessing pipeline of OpenFace is similar to our proposed tool, LabelFace, where both of them are extracting the facial features before classification. OpenFace shows good performance on AU detection of adults compared to deep learning based approaches.

In this paper, we provide a comparison between LabelFace and OpenFace in terms of AU detection performance on children's data.

3 LabelFace: A Facial Expression Processing Tool

Deep learning applications require a significant amount of data. Labeling process of the data takes long time and is very challenging especially in action unit recognition. To ease the labeling process of image and video data, here we present a labeling tool for facial expression, LabelFace. The tool's primary abilities are data labeling, image processing, and video processing.

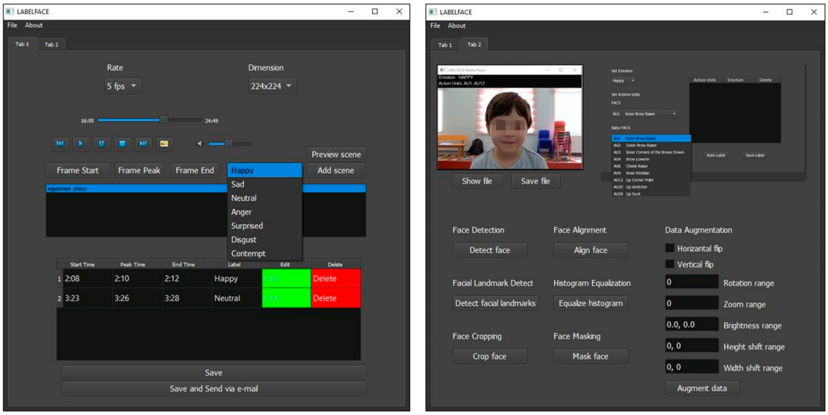
First of all, the tool offers emotion detector models for assisting the user with the labeling process. The tool automatically labels the frames and time frames from a video, and it provides editing options for users to apply changes on labels. As a labeling tool, it lets the user select time frames from a video and label them in different emotion categories (Fig. 2). This process proceeds by specifying the time of the beginning of the expression (Frame Start), the peak time of the expression (Frame Peak), and the end of the expression (Frame End). For the FER on adults, the action units are served according to FACS. For the FER on children, some of the AU's from FACS are served since children do not express all of the AU's presented in FACS.

As an image processing tool, LabelFace provides several image processing methods that users can apply on frames. These processing operations involve face detection, face alignment, facial landmark detection, data augmentation, histogram equalization, face cropping, and face masking.

[1] https://www.kairos.com/.
[2] https://skybiometry.com/.
[3] https://findface.pro/en/.
[4] https://www.noldus.com/facereader.

As a video processing tool, the user can extract frames between defined time intervals. The user can specify the rate of extraction in frame per second (fps) and the extracted frame size. Also, video size, dimension, and view property changes are possible.

Besides the user can apply all these processing operations separately to a dataset, we created a pipeline to apply operations to a dataset directly and get final ready frames or videos to feed into classifiers.



Rate: 5 fps

Dimension: 224x224

	Start Time	Peak Time	End Time	Label	Edit	Delete
1	2:08	2:10	2:12	Happy	Edit	Delete
2	3:23	3:26	3:28	Neutral	Edit	Delete

Emotion : HAPPY
Action Units: AU1, AU12

Set Emotion: Happy

FACS: AU1 Inner Brow Raiser

□ Horizantal flip
□ Vertical flip
0 Rotation range
0 Zoom range
0.0, 0.0 Brightness range
0, 0 Height shift range
0, 0 Width shift range

Fig. 2. LabelFace user interface

Most importantly, the tool provides emotion-assistant models capable of recognizing emotions and detecting action units in children. This capability lets the tool to label the data automatically and assist the user with the labeling process. On the other hand, the emotion assistant models are distributed and can be used for any research purposes.

Besides all of its functions, the tool is a simple application to listen to and watch videos and audio files, and visualize frames with the built-in playlist. It uses QtMultimedia and QtMultimediaWidgets to handle playback and manage the playlist. The main interface offers a playlist window in which users can drag and drop media files to be played. Standard media controls are provided along with a timeline scrub widget and volume control.

Last but not least, we are working on adding deep learning-based FER methods for the user to train their models with ease in the tool interface directly. By adding state of the art deep learning architectures to our tool, we aim to provide researchers a framework in which they can upload their data, and train their private models for specific applications. The tool will offer different deep learning architectures for different research purposes.

4 Test Setup, Implementation and Experiments

The whole experiment setup for facial emotion recognition, including data preparation and data processing, can be seen in Fig. 3. Although we have gathered data from children and applied processing methods using LabelFace, the low quantity of the data has led us to use well-known FER data sets to train deep learning models. A part of the gathered data is then used for testing to see if the trained models are reliable and generalized well.

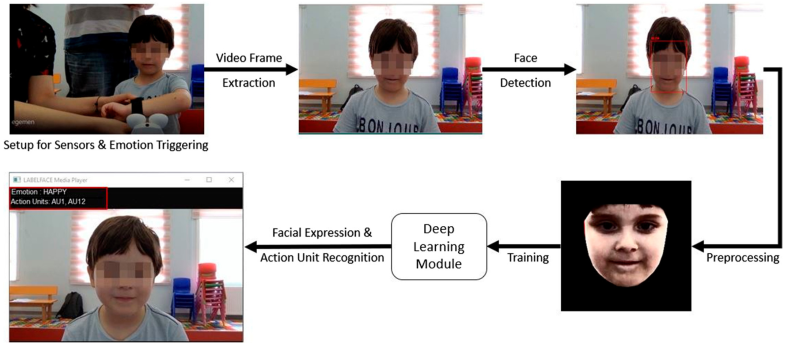


Fig. 3. Facial expression recognition pipeline, Emotion: Happy, Detected action units: AU1, AU12

4.1 Datasets

The Child Affective Facial Expression (CAFE): CAFE dataset is comprised of 2- to 8-year old children poses who have various appearances because of their ethnicity. Children pose seven basic emotions in the dataset: sadness, happiness, surprise, anger, disgust, fear, and neutral. The dataset is also extended with the different appearances of the same emotion by showing the emotion with either open mouth and closed mouth. The entire CAFE set consists of 1192 images, which shows a large variety of emotions to let researchers work on specific problems of the area [22].

The Dartmouth Database of Children's Faces (DDCF): DDCF dataset includes eight different emotion poses from 40 male and 40 female children who are between 6 and 16-year-old. Each child has posed images from different camera angles for each emotion under different light conditions. The subjects are dressed in black hats and clothes to decrease variations between images. Children from 6- to 9-year old are intentionally picked from the DDCF dataset to train emotion detector for children since older children show similarities with adults regarding facial expressions [5].

Our Dataset: We collected non-posed videos from 5- to 9- year old, 16 female and 19 male hearing-impaired children. During data collection, 18 children were asked to watch different video content from well-known animated movies to trigger different emotions. The rest of the children attended a CRI study, where they are asked to take a test on a tablet while being assisted by a Pepper humanoid robot. The children take two tests, 1 with only tablet and the second test with a tablet and robot. In both cases, the children's reactions are video recorded. We extracted frames with significant emotional intensity levels from video recordings of the study to be used in our models' testing phase. The extracted images are then labeled with 12 action units shown in Table 1. We obtained 198 emotionally intense frames from the videos, and these frames are augmented with shifting, flipping, rotating operations. Finally, we got 1422 frames which are then separated into training, validation and testing for fine-tune experiments.

4.2 Proposed Method

We used the aforementioned datasets to train AU detection models. CAFE and DDCF [5] are only available datasets with emotion labels. We used peak frames from the videos of the children with hearing impairments for non-posed data evaluation, and DDCF dataset for posed data evaluation. Thus, we labeled these datasets according to FACS and baby FACS [9]. Target AU's for the children, their distribution in the posed DDCF dataset and in our non-posed dataset can be seen in Table 1. We first extracted and labeled 12 AUs which are mostly found in children's facial expressions. Afterwards, for a fair comparison with OpenFace, we included only nine common action units, which can be recognized in both tools: AU1, AU4, AU5, AU6, AU9, AU10, AU12, AU15, and AU17.

Table 1. AUs coded by manual FACS coders for the peak frames and their distributions in the children datasets (* belongs to baby FACS [9], † are excluded for comparison)

AU	Description	CAFE (344)	DDCF (180)	Ours (198)
0	No Action Unit	71	38	24
1	Inner Brow Raiser	92	47	22
3*†	Brows Drawn Together	71	18	16
4	Brow Lowerer	58	28	7
5	Upper Lip Raiser	61	33	6
6	Cheek Raiser	51	25	65
9	Nose Wrinkler	55	15	10
10	Upper Lip Raiser	55	6	3
12	Lip Corner Puller	78	42	88
15	Lip Corner Depressor	41	23	13
17	Chin Raiser	44	39	27
24†	Lip Pressor	29	51	54
27†	Mouth Stretch	86	36	6

We utilized preprocessing methods to remove variations between subject frames. The data preparation process involves face detection, face alignment, data augmentation, histogram equalization, face cropping, and face masking. As a first step, we detect the faces to remove irrelevant parts from the input frame. For this aim, a CNN based face detector model from dlib library [15] is used. Dlib is a powerful tool to extract facial landmark points from a facial image, that will be used for further processing operations. The next step, face alignment is the geometric adjustment of the face according to facial landmark points based on scale, translation and rotation. Histogram equalization is another processing method that aims to improve contrast in images by manipulating image pixel intensities. After processing the input frames with the aforementioned methods, we crop the facial image and we use facial landmark points to overlay the face with a mask. The masked frames are then augmented with the randomized flipping, rotating, shearing and cropping operations. The augmented frames are finally resized to 224 × 224 RGB images. The example of the original images and the final masked images belonging to subjects in the datasets can be seen in Fig. 4.

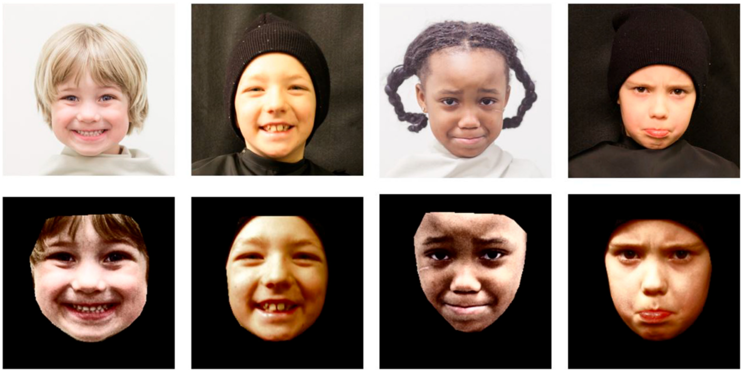


Fig. 4. Original images vs. masked images of subjects belonging to CAFE, DDCF datasets respectively

After data processing steps, we obtained 2335 CAFE frames, and 1190 DDCF frames. This quantity was not sufficient for efficiently training a deep learning model; thus, we use transfer learning which involves a pre-trained model as a baseline model for specialized purpose training [25].

The pre-trained model is trained on a large amount of facial data. Instead of training from scratch, keeping the pre-trained model as a baseline and fine-tuning it with a small amount of data can give better results. This process gives a good initialization point for the task.

As a baseline model, VGG-FACE architecture is utilized. VGG-FACE is a deep convolutional network that is modified and used for different face recognition tasks. The model accepts 224 × 224 RGB images as inputs and outputs class probabilities with softmax function [26]. VGG-FACE model is trained with 2.6 million face images, and integrates significantly high amount of facial data which provides a robust model for face detection. Compared to other architectures, ResNet [10] and VGG-FACE show the most successful performances [21]. VGG-FACE is also one of the most studied architectures in the literature [2,7,30]. Based on these motivations and the similar characteristics of our own data set obtained from real world setup, we extended VGG-FACE with customized top layers to fulfill the requirements of our system at best. Our model is an extended version of the VGG-FACE, which is based on VGG-Very-Deep-16 CNN architecture [29]. The model architecture can be seen in Fig. 5.

Table 2. Performance analysis based on the recognized AU number

Training set	Test set	Precision	Recall	F1 score	Number of AUs
CAFE	DDCF	75.3	66.0	68.1	9
CAFE	DDCF	80.6	58.4	60.5	12
OpenFace 2.0	DDCF	18.1	39.1	24.2	9
CAFE	Our dataset	61.1	45.6	47.1	9
CAFE	Our dataset	67.6	34.0	41.0	12
CAFE + DDCF	Our dataset	61.9	46.5	50.7	9
CAFE + DDCF	Our dataset	67.0	39.0	48.0	12
OpenFace 2.0	Our dataset	32.0	52.1	36.2	9

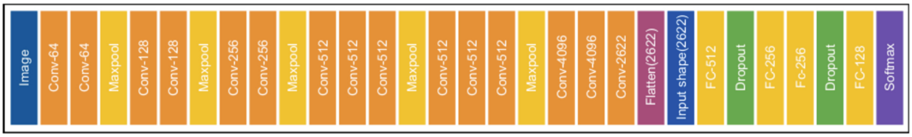


Fig. 5. VGG-FACE architecture with custom top player model

VGG-FACE is used as an encoder network which extracts features from training data. As one of the contributions of this study, for transfer learning, fully connected layers of VGG-FACE is changed with convolutional layers and a new top model with fully connected layers placed for action unit recognition. The extracted features are used for training by the custom top layers. The new base model includes 16 convolutional layers, five pooling layers, and the top layer includes four fully connected layers. Dropout layers are added to prevent overfitting. The final layer of the network has 12 nodes, which corresponds to 12 AU's. Training parameters are kept the same with the general purpose of use (RELU activation, batch size (16), strides (1,1), Adam optimizer, mean squared error (MSE), learning rate (0.0001), momentum (0.9)). LabelFace is implemented using Python programming language, and the proposed deep learning models are trained with Python, Keras-Tensorflow libraries.

5 Test Results and Discussions

The experiments are conducted on several different setups. Only the VGG-FACE baseline model kept the same, and different configurations for the top layer are applied. The evaluation of the models has been done with cross-database evaluation; precision, recall, and F1-scores are used as evaluation metrics [16]. The evaluation results of the models are summarized in Table 2.

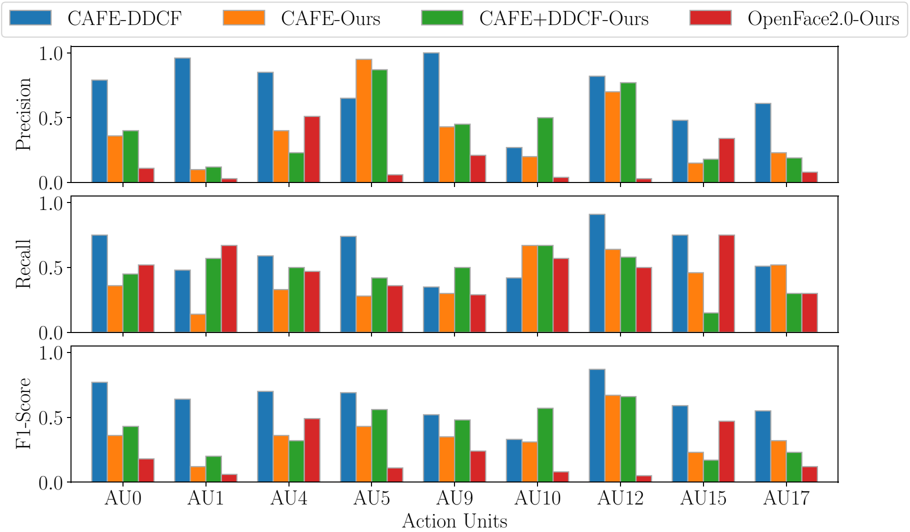


Fig. 6. Precision, recall and F1 Scores for each AU

For the comparison with OpenFace on posed-children data, we first used 9 common AUs. As a first step we fed OpenFace with the DDCF dataset. OpenFace achieved %18.1 precision, %39.1 recall, and %24.2 F1-score on the evaluation of 175 DDCF frames. At the next step, we tested OpenFace with our non-posed children dataset to see the tool's generalization performance. OpenFace achieved %32 precision, %52.1 recall, and %36.2 F1-score on the evaluation of 198 non-posed frames. The precision, recall and F1-score distributions of 9 AUs are showed in the Fig. 6.

For children's AU detector model training, CAFE is chosen as the primary dataset because it contains relatively large amount of data that is balanced in terms of AUs. We trained the top layer for 500 epochs and observed the best performance around 250 epochs. The evaluations of the trained model with DDCF and our datasets showed that the CAFE model is successful in terms of generalization and reliability. The model shows more success in posed-data (F1-score of %68.1 with testing DDCF) which is not surprising since CAFE is a posed-dataset. The model achieves F1-score of %47.1 on real-world non-posed data.

The next step is fine-tuning of the model with another dataset to observe whether the fine-tuned model generalizes well. Thus, in the last experiment, the previously trained CAFE model is fine-tuned with the DDCF dataset to examine the new model's performance with our dataset. The model is fine-tuned for 100 epochs, and a better evaluation result is observed with non-posed data, which also surpasses the original CAFE model by achieving F1 score of %50.7. When we compare the results of the model trained with CAFE and tested with DDCF, which has the highest F1 score, the results show that the models tested with our dataset have a lower F1 score. The differences in model performances are mostly caused by the fact that, the other datasets are composed of posed frames, i.e., the children express exaggerated emotions, whereas in our dataset, we take non-posed images with more natural expressions of emotions, at wild. The non-posed dataset is taken during a child-robot interaction game featuring Audiology test, as stated above. The children were focused on the robot, and showed less arousal in the emotions during the interaction which also decreases the difference between their facial expression in different emotions.

We also extended the AU detection capability by adding AU3, AU24, and AU27, that are among the encountered AUs in training sets. The precision, recall and F1-score distributions of the 12 AUs are displayed in the Fig. 7.

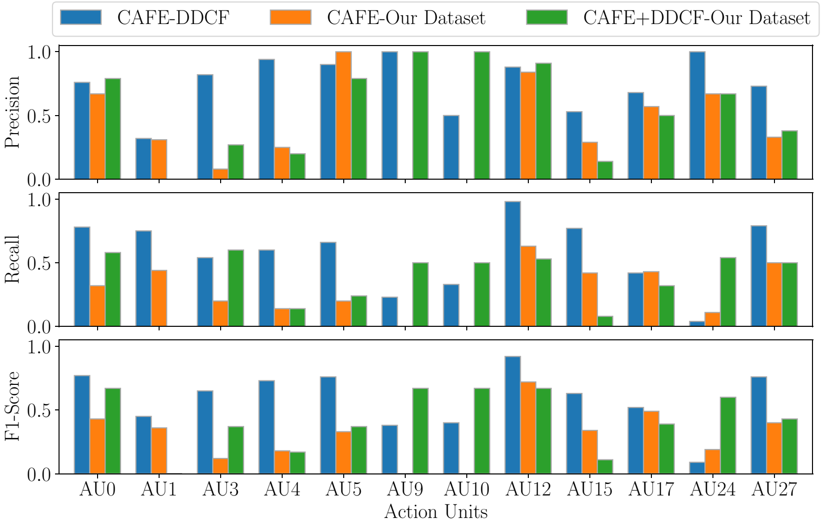


Fig. 7. Precision, recall and F1 Scores for 12 AUs

Posed data evaluation (CAFE - DDCF) results show that AU0 (absence of action units), AU3, AU4, AU5, AU9, AU12, AU17, AU24 and AU27 precision scores are satisfying, except AU1. We observed that the intensity values of AU1 samples in our dataset is low compared to other AUs, which might be the reason of the low detection performance. When we look at the recall scores, we observe the drastic change in AU24 compared to precision, which might be caused by the fact that AU24 is represented by less number of training samples in CAFE.

For the non-posed data evaluation (CAFE - Our Dataset and CAFE+DDCF - Our Dataset), the unbalanced distribution of the test set, especially the low number of frames representing some of the action units, makes it difficult to evaluate action unit recognition results. Overall, the performance of fine-tuned model is satisfactory with the most occurred AUs which are AU0, AU5, AU9, AU10, AU12, and AU17.

In summary, it is observed that the classifier's performance increases by preprocessing of the datasets, which is also observed in previous studies in the literature [19]. On the other hand, locating the specially designed layers which are trained using the datasets, on the top of the pre-trained VGG-FACE, has increased the robustness of the classifiers. Also, it is observed that, when a classifier which is trained by one dataset is fine-tuned by another dataset, its performance in recognition of children's emotions has increased. The performance differences between LabelFace and OpenFace on children's AU detection is due to a couple of reasons. First of all, OpenFace aims to respond to all age scales of the FER problem. Therefore, OpenFace AU detector models are trained with datasets that include both children and adult facial images. In other words, OpenFace uses the same AU detector models for all age ranges. But in our study with children, using only children datasets seems to improve the performance of AU detectors. Thus, it can be derived that AU detection for children and adults should be considered within different contexts. Another reason for the difference in the performance is the approaches used for AU detection. Using transfer learning with neural networks provides better results than a linear SVM classifier in complex tasks such as AU detection.

6 Conclusion and Future Directions

This study aims to build an emotion recognition system including a tool that handles data labeling, data processing, and model training with ease of use. The system is designed and developed to be used with an affectively aware humanoid robot for children with hearing impairments.

Facial emotion recognition is one of the most frequently used methods in emotion recognition studies. One of the most challenging phases in this studies is data labeling and processing. Although there are numerous tools for labeling of visual data, there are not sufficient tools for emotion recognition, which is used for different user groups with different specifications, such as children.

In this paper, we presented an emotion labeling and facial expression/action unit processing tool, LabelFace. LabelFace provides data labeling, data processing, emotion assistant functions to the user. The tool serves these functions either separately or in a pipeline manner. The prepared pipeline includes visual data processing operations to prepare user data for model training. The tool provides AU detection models to help users label their data independent from the data size.

Pre-processed datasets are fed to deep learning models using transfer learning and fine-tuning of cross-datasets. The test evaluations of the different setups show that, when VGG-FACE is used together with the specially designed layers and transfer learning approaches, the models perform more successfully in children's facial expressions. In our approach, multi-label action unit detection models are used. As of our knowledge, our proposed children AU detector model is the first model which targets 5- to 9- year old children. The model trained with well-known posed datasets (CAFE, DDCF) and tested with real-world nonposed dataset collected from our children-robot interaction studies with hearing-impaired children.

Compared to a well-known open source facial expression processing tool, OpenFace 2.0, LabelFace performed more successfully in the detection of the AUs, which are commonly observed in both models. The most succesfully detected AUs by our models are AU0, AU5, AU9, AU10, AU12, and AU17, which are the components of neutral, positive and negative emotions.

Unlike adults, children do not show all AUs of the emotions, especially when they are not posing, therefore using affective tools such as LabelFace for the detection of AUs will help healthcare workers, therapists and caregivers in the detection of children's emotion and stress.

LabelFace is not publicly released yet, but we are working on the release of the first version in the first half of the next year. Also, we are working on the extension of tool capabilities by introducing other deep learning models, that do not only aims children, but also adults. We aim to investigate the most frequent AUs in children, and plan to minimize the number of required AUs to recognize emotions, stress and attention in children in the wild, which will be very beneficial in digital healthcare systems designed for children, especially children with special needs. We also intend to enhance the tool capabilities by introducing auto-labeling the data of adults. For this aim, we investigate diverse adult datasets along with the appropriate deep learning models.

Acknowledgment. This study is supported by the Scientific and Technological Research Council of Turkey (TUBITAK), RoboRehab project, under contract no 118E214.

References

1. Al-agha, L.S.A., Saleh, P.H.H., Ghani, P.R.F.: Geometric-based feature extraction and classification for emotion expressions of 3D video film. J. Adv. Inf. Technol. **8**(2), 74–79 (2017)
2. Albiero, V., Bellon, O., Silva, L.: Multi-label action unit detection on multiple head poses with dynamic region learning. In: 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, pp. 2037–2041. IEEE, October 2018. https://doi.org/10.1109/ICIP.2018.8451267
3. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), Xi'an, China, pp. 59–66. IEEE (2018)

4. Breuer, R., Kimmel, R.: A deep learning perspective on the origin of facial expressions. arXiv preprint arXiv:1705.01842 (2017)
5. Dalrymple, K.A., Gomez, J., Duchaine, B.: The Dartmouth database of children's faces: acquisition and validation of a new face stimulus set. PLoS ONE **8**(11), e79131 (2013)
6. Ekman, P., Friesen, W.V.: Constants across cultures in the face and emotion. J. Pers. Soc. Psychol. **17**(2), 124 (1971)
7. Ertugrul, I.O., Yang, L., Jeni, L.A., Cohn, J.F.: D-pattnet: dynamic patch-attentive deep network for action unit detection. Front. Comput. Sci. **1**, 11 (2019)
8. Friesen, W.V., Ekman, P.: Facial action coding system: a technique for the measurement of facial movement. Palo Alto vol. 3 (1978)
9. Hammal, Z., Chu, W.S., Cohn, J.F., Heike, C., Speltz, M.L.: Automatic action unit detection in infants using convolutional neural network. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, TX, USA, pp. 216–221. IEEE (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 770–778. IEEE, June 2016. https://doi.org/10.1109/CVPR.2016.90
11. Huang, Y., Yang, J., Liao, P., Pan, J.: Fusion of facial expressions and EEG for multimodal emotion recognition. Comput. Intell. Neurosci. **2017**, 1–8 (2017)
12. Huang, Y., Chen, F., Lv, S., Wang, X.: Facial expression recognition: a survey. Symmetry **11**, 1189 (2019). https://doi.org/10.3390/sym11101189
13. Jain, N., Kumar, S., Kumar, A., Shamsolmoali, P., Zareapoor, M.: Hybrid deep neural networks for face emotion recognition. Pattern Recogn. Lett. **115**, 101–106 (2018)
14. Jiang, Y., Li, W., Hossain, M.S., Chen, M., Alelaiwi, A., Al-Hammadi, M.: A snapshot research and implementation of multimodal information fusion for data-driven emotion recognition. Inf. Fusion **53**, 209–221 (2020)
15. King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
16. Ko, B.: A brief review of facial emotion recognition based on visual information. Sensors **18**(2), 401 (2018)
17. LeCun, Y., Haffner, P., Bottou, L., Bengio, Y.: Object recognition with gradient-based learning. Shape, Contour and Grouping in Computer Vision. LNCS, vol. 1681, pp. 319–345. Springer, Heidelberg (1999). https://doi.org/10.1007/3-540-46805-6_19
18. Leppänen, J.M., Nelson, C.A.: The development and neural bases of facial emotion recognition. In: Advances in Child Development and Behavior, vol. 34, pp. 207–246. Elsevier (2006)
19. Li, S., Deng, W.: Deep facial expression recognition: a survey. IEEE Trans. Affective Comput. 1 (2020)
20. Lim, N.: Cultural differences in emotion: differences in emotional arousal level between the east and the west. Integrative Medicine Research **5**(2), 105–109 (2016). https://doi.org/10.1016/j.imr.2016.03.004. http://www.sciencedirect.com/science/article/pii/S2213422016300191
21. Lim, Y.K., Liao, Z., Petridis, S., Pantic, M.: Transfer learning for action unit recognition. ArXiv abs/1807.07556 (2018)
22. LoBue, V., Thrasher, C.: The child affective facial expression (CAFE) set: validity and reliability from untrained adults. Front. Psychol. **5**, 1532 (2015)

23. Mollahosseini, A., Chan, D., Mahoor, M.H.: Going deeper in facial expression recognition using deep neural networks. In: 2016 IEEE Winter conference on applications of computer vision (WACV), Lake Placid, NY, USA, pp. 1–10. IEEE (2016)
24. Oster, H.: Baby FACS: Facial action coding system for infants and young children. Unpublished monograph and coding manual (2000)
25. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010). https://doi.org/10.1109/TKDE.2009.191
26. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: Xie, X., Jones, M.W., Tam, G.K.L. (eds.) Proceedings of the British Machine Vision Conference (BMVC). pp. 41.1–41.12. BMVA Press, Swanse, September 2015. https://doi.org/10.5244/C.29.41. https://dx.doi.org/10.5244/C.29.41
27. Ranganathan, H., Chakraborty, S., Panchanathan, S.: Multimodal emotion recognition using deep learning architectures. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), , Lake Placid, NY, USA, pp. 1–9. IEEE (2016). https://doi.org/10.1109/WACV.2016.7477679
28. RoboRehab: Assistive audiology rehabilitation robot. https://roborehab.itu.edu.tr/. Accessed 21 Oct 2020
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
30. Tang, C., et al.: View-independent facial action unit detection. In: 2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017), Washington, DC, USA, pp. 878–882. IEEE, May 2017. https://doi.org/10.1109/FG.2017.113
31. Yu, Z., Liu, G., Liu, Q., Deng, J.: Spatio-temporal convolutional features with nested LSTM for facial expression recognition. Neurocomputing **317**, 50–57 (2018)