Gdańsk University of Technology
Faculty of Electronics, Telecommunications and Informatics

Technical report 01/2022

Analyzing the emotion recognition capabilities based on video recordings of the faces of children on the autism spectrum while interacting with a social robot

Agata Kołakowska, Jan Kowalina, Agnieszka Landowska, Michał Wróbel, Ihar Uzun

This publication was supported in part by the Erasmus Plus project of European Commission: EMBOA, Affective loop in Socially Assistive Robotics as an intervention tool for children with autism, contract no 2019-1-PL01- KA203-065096. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein. This publication is distributed free of charge.

1. Introduction

The aim of the analytical work was to identify problems with emotion recognition from video recordings of the faces of children on the autism spectrum during interaction with a social robot, with a view to developing guidelines for future research of this type.

Automatic emotion recognition was carried out with Noldus FaceReader 8 software. The processing of video files with recorded interactions of children with the robot resulted in output files with recognised emotions. In order to determine the capability of emotion recognition from video recordings, in a first step the rate of emotion detection in each file was determined. Each frame was marked as:

- **FIND_FAILED** – could find the face
- **FIT_FAILED** – could not fit the face model
- **DETECTED** – emotion was detected

All videos were then manually reviewed and problems were identified that may have affected the level of emotion detection. Finally, the most common problems recurring in recordings with a low rate of emotion recognition are grouped and described

2. Facial expression recognition

Information from various sources may be analyzed to infer one's emotional state, i.e. facial expression, voice, physiological signals, behavioral patterns. Among them, facial expressions, which is the natural nonverbal source of information utilized by humans to get to know others' emotions, seem to be the most effective in the automatic recognition as well. Therefore numerous method have been developed for years, achieving higher and higher accuracies.

Emotion recognition methods based on face expression analysis belong to one of two main categories. The first one are traditional methods based on handcrafted features. The second approach lets a deep neural network extract features. Among these solutions, convolutional neural networks became the most popular ones.

2.1. Face detection

No matter of the approach applied, face detection is the first step performed before the analysis of expression may begin. The first algorithm, that gave satisfying results in real-world conditions was proposed by Viola & Jones (2004). Their method is based on a number of Haar-like features as the image representation. These features represent differences of pixel intensities within rectangular areas. There are several types of them and they are evaluated at different locations on an image. AdaBoost cascade classifier is applied to select the best features for the given task and to train a classifier able to detect faces on the basis of these characteristics. The method both in its original form and with some modifications is commonly applied to detect faces. The algorithm performs very well for frontal face images. Sometimes it is used as the first detection step filtering most of non-face images and followed by another detector. Neural networks have also been widely applied in the task of face detection (Zafeiriou, 2015). Deep neural networks have turned to be especially successful in this area in recent times. An example of a detector of this type has been presented by Zhang & Zhang (2014), who trained a multitask deep convolutional neural network (DCNN) for multiview face detection task. The image is first passed through a cascade-based multiview face detector. If it is not rejected then it is preprocessed and sent to DCNN to make up the final decision. The role of the network is to improve the effect of a simpler detector. The dataset used to train the network contained images with frontal, half profile and profile faces. The model is not only trained to detect face, but also the facial pose and the location of seven landmarks.

2.2. Feature extraction

The second stage of face expression recognition process is feature extraction. In the case of traditional approach based on handcrafted features two types of face descriptors may be extracted, either appearance or geometric ones (Ko, 2018). Appearance features describe the texture of facial image, whereas geometric ones focus on the shape of face and its elements.

Among appearance features, local binary patterns (LBP), histograms of oriented gradients (HoG) or scale invariant feature transform (SIFT) are one the most popular. LBP are good texture features representing the texture locally and forming their occurrence histogram (Ojala et al., 2002). LBP are not only rotation invariant, robust to illumination variations, but also robust to changes in gray scale. It is calculated by comparing a pixel's value with its eight neighbours values and assigning 0 when the value is greater or 1 otherwise. In this way a pixel is encoded with an 8-bit number. Then a histogram of these numbers occurrences is created. The histogram is a 256-dimensional vector, which is a good texture descriptor. Another method, called histograms of oriented gradients (HoG) employ occurrences of gradient orientation in localized portions of an image (Dalal, 2005). For each pixel the magnitude and direction of gradient is calculated. Then the image is divided into cells and a 9-bin histogram of gradients for each cell is calculated. The histogram of cells are combined in a specific way to form a single vector. HoG features are invariant to geometric and photometric transformations, except for object orientation. Another feature generation method is scale invariant feature transform (SIFT) introduced by Lowe (1999). It lets find locations invariant to translation, scaling, rotation and minimally affected by noise and small distorsions. Then it finds a representation for these regions. The obtained features are local, so they are also robust to occlusion. Another popular method used to extract features of face images are Gabor filters. A number of Gabor filters corresponding to different resolutions and orientations are usually applied and a set of features is extracted from the filtered images (Lyons, 1998). Subspace projection techniques have been also widely applied for face representation as appearance features, some of them in the task of facial expression analysis, e.g. non-negative matrix factorization (Ale et al., 2015).

Some of the methods, e.g. Gabor transform, lead to high number of features, so dimensionality reduction is often performed either by performing feature selection or by transforming data to a new space of lower dimension, e.g. by applying principal component analysis or linear discriminant analysis (Sariyanidi et al., 2015).

Appearance features may be extracted either for the whole image or for selected regions. The region may be defined by applying a grid to the image, however best results are obtained for specific local regions, e.g. found automatically using landmark localisation methods. Feature vectors are then extracted for selected regions depending on a given task. For example mouth and eyes regions are one of those carrying the most discriminating information on face expression (Ghimire et al., 2017).

The other type of characteristics are geometric features, which are based on the location of characteristic points. Examples of features of this type are distances or angles between the landmark points or normalized central moments calculated for the points. There are various methods of landmark detection algorithms. The classic Active Appearance Model proposed by Cootes et al. (2001) is widely used. It creates a statistical model on the basis of training set, which consists of images with coordinates of landmark points. Then the model is matched with new images. This method lets match shape and texture simultaneously in contrast to a prior idea of active shape models (ASM) (Cootes et al., 1995).

Another method used for landmark location is based on a mixture of trees (Zhu & Ramanan, 2012) and it combines face detection and landmark localization. The landmarks are modeled as tree parts taken from a common set. Trees are applied to model the topological changes between different views of an object. The model turned out to be effective in capturing global deformations also for single viewpoints, which occur while face expression changes.

The landmark points may be used not only to calculate some characteristic features describing shape, but they may also be tracked while analyzing video sequences. In such dynamic scenario, the displacements of selected points between subsequent frames may be used to calculate dynamic features describing for example face expression (Ko, 2018). Some researchers implement both landmark and appearance features. A common approach is to localize key points and then describe regions around these points using appearance parameters.

2.3. Classification

The feature extraction phase is then followed by classification, which may be performed using various models, e.g. SVM, random forest, AdaBoost and many others. Nowadays most methods used to recognize emotions on the basis of face expression apply deep neural networks. The main advantage of applying neural networks is the fact that the network is able to extract important features of the input data, depending on a given training set. Thus the traditional handcrafted features do not have to be implemented. The networks usually used are convolutional networks (CNN), which are especially suitable for processing 2D images. A set of convolution and pooling layers constitute a feature extractor. They are followed by fully connected layers which play role of a classifier outputting the final decision. Sometimes the network is used only to extract features and then another model is used as a classifier, e.g. SVM or AdaBoost.

2.4. Other solutions based on deep neural networks

Although neural networks are able to extract features, they are sometimes combined with traditional feature extraction procedures to incorporate well known good image descriptors. For example Levi & Hassner (2015) applied local binary patterns (LBP) to encode images. Then they mapped the images represented in this way to a 3D space using multidimensional scaling. The LBP were calculated with different parameter settings leading to three representations. These representations and the original RGB images constituted input data for an ensemble of CNNs of different architectures. Each combination of representation and architecture was estimated and it turned out than LGB representations usually outperformed the original ones. Another example of incorporating traditional features into neural networks was presented by Zhang et al. (2016), who extracted scale-invariant features (SIFT) and then used them as input to a deep neural network. The network was trained to find an optimal set of discriminative features for recognizing face expression in the case of varying facial views. It is also worth taking into account the landmark locations while designing neural network input, because it may reduce the problem of variations in scale and rotation (Li et al., 2020).

Convolutional neural networks of various architectures are the most common ones, due to their ability to take into account the spatial layout of image pixels. However, other networks are also used. For example Usman (2017) applied an autoencoder to reduce the dimensionality of the extracted HoG features, then he trained the SVM classifier to recognize emotions.

Researchers investigate various interesting network architectures trying to enhance the quality of facial expression recognition. One of the latest ideas is to apply a convolutional network with the attention mechanism to make the network focus on the essential regions of the face (Minaee, 2021). This led to a network with less than 10 layers but able to achieve results as good as much deeper convolutional networks usually trained for this task.

Some studies suggest that applying multitask learning may be beneficial for face expression recognition. Networks are trained on the basis of data sets from different sources and labeled in different ways. Pons (2018) described a solution based on a network which has been trained both to recognize face expression and to detect action units. Specially defined selective cross-entropy loss function enabled sharing the whole network among the two tasks and images from two databases, even if an input image was not labeled for both tasks.

Other types of neural networks applied in face expression recognition are deep belief networks, deep autoencoders, recurrent neural networks and generative adversarial networks (Li & Deng, 2020). An example solution based both on convolutional and recurrent networks was presented by Kim et al. (2019), who first applied a convolutional network to learn spatial characteristics of facial images. Then some temporal characteristics of this representation were learned using a long short-term memory.

The amount of data needed to train a complex model, such as deep neural networks, is a substantial problem. There are some publicly available face expression recognition data sets, but their size is often not sufficient enough to avoid the problem of overfitting. One of the ideas applied to overcome this issue is to pre-train a network using another data set, e.g. prepared for face recognition, and then fine-tune it using the face expression images. It is also possible to apply one of available models and only fine-tune that model using own data as for example in (Dong et al., 2021), where the well known VGG16 network pre-trained on ImageNet data set was then fine-tuned using facial expression data. The approach let achieve higher recognition accuracy than after training only on the basis of expression images.

Another idea, commonly implemented to cope with the problem of insufficient amounts of data or imbalance in class distribution, is data augmentation. Various operations are performed to generate additional training examples, e.g. horizontal flip, rotation, scaling, adding noise, changing contrast. Sometimes syntetic images are also generated using a specially designed model (Li & Deng, 2020).

Although the proposed solutions let achieve better and better accuracies, the analytical work presented in this report reveals that real-life environment, where occlusions or variations in pose and illumination may occur, or where interpersonal variations, e.g. in expression intensity, are inevitable, still makes facial expression recognition a challenging task.

3. FaceReader

To perform the experiments, FaceReader software was used. It allows to classify emotions on the basis of facial expressions. Several emotional states may be recognized using FaceReader, i.e. six Ekman's basic emotional states (happy, sad, angry, surprised, scared, disgusted), contempt and neutral state.

Due to the fact that an expression usually results from a mixture of emotions, the software analyzes the face taking into account all possible states. Each expression is assigned a value from 0 to 1, depending on its intensity.

The recognition process consists of the following three stages:

1. Detecting the position of the face in the image performed using a deep learning algorithm (Zafeiriou et al., 2015).

2. Modeling the face by finding 468 key points and then reducing the dimension of this representation by performing principal component analysis. According to Noldus documentation the landmark localisation was performed by applying a deep neural network presented by Bulat et al. (2017). The method described in the mentioned paper is based on one of the state-of-the-art architectures for human pose estimation. Three different networks have been trained. The first one is trained to convert 2D landmark annotations to 3D and was designed to create a large-scale 3D face alignment annotations dataset. The other two networks were trained to find 2D and 3D landmark locations respectively.

3. Classification using a neural network trained on the basis of over 20000 images annotated by human experts. The network has been trained to recognize emotional states and a set of face action units. The Noldus documentation refers to the network architecture presented by Gudi et al.

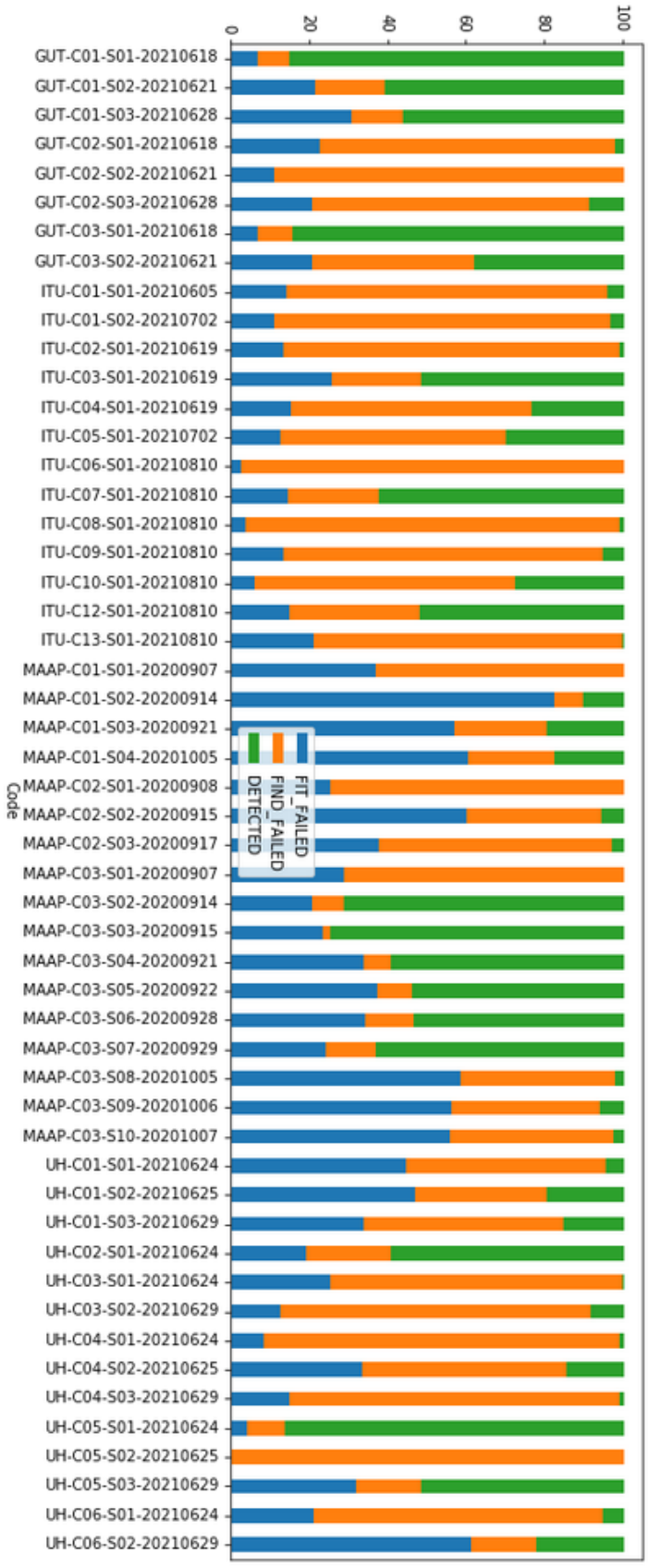
(2015). It consists of three convolutional layers. The first one is followed by a pooling layer, whereas the third one is followed by a fully connected layer. The output layer consists of a number of neurons, one per each class. The ReLu activation function has been applied in all neurons.

Apart from the above mentioned emotional states, FaceReader also analyzes valence and arousal. Valence indicates whether the emotion is positive or negative. It is calculated as a difference between the intensity of happiness and the intensity of one of negative states (sad, angry, scared, disgust), which is assigned the highest intensity. Arousal indicates whether the person is active or not. It is calculated on the basis of a set of selected action units.

The recognition of 20 action units, i.e. muscle groups responsible for facial expressions, also adds valuable information. Some emotional states, which do not belong to the basic set of emotions recognized by the tool, may be inferred from the estimated intensities of selected action units.

According to the software documentation the performance of the tool has been validated using the Amsterdam Dynamic Facial Expression Set (ADFES), which contains images of posed eight emotional face expressions, achieving accuracy of 100% for all emotional states except for sadness (95,8%).

4. Emotion recognition rates



Code	FIT_FAILED	FIND_FAILED	DETECTED
UH-C05-S01-20210624	4,01%	9,38%	86,60%
GUT-C01-S01-20210618	6,42%	8,38%	85,20%
GUT-C03-S01-20210618	6,46%	9,05%	84,49%
MAAP-C03-S03-20200915	23,02%	1,97%	75,00%
MAAP-C03-S02-20200914	20,39%	8,15%	71,46%
MAAP-C03-S07-20200929	21,79%	12,25%	65,97%
ITU-C07-S01-20210810	14,33%	23,23%	62,44%
GUT-C01-S02-20210621	21,21%	17,62%	61,17%
UH-C02-S01-20210624	18,81%	21,73%	59,46%
MAAP-C03-S04-20200921	33,73%	6,94%	59,33%
GUT-C01-S03-20210628	30,32%	13,32%	56,36%
MAAP-C03-S05-20200922	36,97%	8,81%	54,22%
MAAP-C03-S06-20200928	33,84%	12,52%	53,64%
ITU-C12-S01-20210810	14,61%	33,30%	52,09%
ITU-C03-S01-20210619	25,41%	22,78%	51,81%
UH-C05-S03-20210629	31,52%	16,85%	51,63%
GUT-C03-S02-20210621	20,37%	41,49%	38,14%
ITU-C05-S01-20210702	12,37%	57,38%	30,25%
ITU-C10-S01-20210810	5,89%	66,21%	27,90%
ITU-C04-S01-20210619	15,08%	61,41%	23,51%
UH-C06-S02-20210629	60,99%	16,48%	22,53%
MAAP-C01-S03-20200921	56,49%	23,72%	19,80%
UH-C01-S02-20210625	46,83%	33,42%	19,74%
MAAP-C01-S04-20201005	60,05%	22,11%	17,84%
UH-C01-S03-20210629	33,53%	51,02%	15,45%
UH-C04-S02-20210625	33,09%	52,35%	14,56%
MAAP-C01-S02-20200914	82,23%	7,34%	10,43%

GUT-C02-S03-20210628	20,28%	71,00%	8,71%
UH-C03-S02-20210629	12,21%	79,35%	8,44%
MAAP-C02-S02-20200915	59,42%	33,88%	6,70%
MAAP-C03-S09-20201006	56,08%	37,77%	6,15%
ITU-C09-S01-20210810	13,12%	81,39%	5,49%
UH-C06-S01-20210624	20,95%	73,65%	5,40%
UH-C01-S01-20210624	44,43%	51,08%	4,49%
ITU-C01-S01-20210605	13,96%	81,89%	4,15%
ITU-C01-S02-20210702	10,68%	85,89%	3,43%
MAAP-C02-S03-20200917	37,26%	59,39%	3,35%
MAAP-C03-S10-20201007	55,64%	41,59%	2,77%
GUT-C02-S01-20210618	22,52%	75,13%	2,35%
MAAP-C03-S08-20201005	58,28%	39,47%	2,24%
ITU-C02-S01-20210619	13,28%	85,42%	1,30%
UH-C04-S01-20210624	8,00%	90,94%	1,06%
ITU-C08-S01-20210810	3,44%	95,52%	1,04%
UH-C04-S03-20210629	14,75%	84,23%	1,01%
ITU-C13-S01-20210810	21,01%	78,51%	0,48%
UH-C03-S01-20210624	25,13%	74,50%	0,37%
GUT-C02-S02-20210621	10,79%	89,05%	0,16%
MAAP-C01-S01-20200907	36,53%	63,32%	0,15%
MAAP-C02-S01-20200908	24,97%	74,92%	0,11%
MAAP-C03-S01-20200907	28,64%	71,27%	0,09%
ITU-C06-S01-20210810	2,15%	97,82%	0,03%
UH-C05-S02-20210625	0,00%	100,00%	0,00%

5. The most common recurring issues for recordings with low emotion recognition rates

1. **Recordings:**
    - low resolution

- overexposed video, very bright light
- long distance from the child's face
- angle of the camera - usually too high in relation to the child's face

2. **Child's appearance:**
   - thick glasses
   - long fringe

3. **Child behavior:**
   - lowers head
   - looks around
   - does not look at the robot
   - covers face with hand
   - plays with hair

6. Acknowledgements

We are very grateful to all children, families, and healthcare professionals for their participation in our study. We want to thank all project partners for data acquisition.

This publication was supported in part by the European Commission's Erasmus Plus project "EMBOA, Affective loop in Socially Assistive Robotics as an intervention tool for children with autism", contract no 2019-1-PL01- KA203-065096.

The European Commission's support for the production of this publication does not constitute an endorsement of the contents which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

7. Bibliography

1. Ali, H. B., Powers, D. M. W., Jia, X., & Zhang, Y. (2015). Extended Non-negative Matrix Factorization for Face and Facial Expression Recognition. *International Journal of Machine Learning and Computing*, *5*(2), pp. 142–147.
2. Bulat, A.; Tzimiropoulos, G. (2017). How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In Proceedings of the IEEE International Conference on Computer Vision 2017, pp. 1021-1030.
3. Cootes, T. F., Taylor, C. J., Cooper, D. H., Graham, J. (1995). Active shape models—their training and application. *Comput. Vis. Image Underst.* 61, 1 (Jan. 1995), pp. 38–59.
4. Cootes, T. F., Edwards, G. J. Taylor, C. J. (2001). Active appearance models, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681-685.

5. Dalal, N. & Triggs, B. (2005). Histograms of oriented gradients for human detection, *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 886-893 vol. 1, doi: 10.1109/CVPR.2005.177.
6. Dong, C., Wang, R., Hang, Y. (2021). Journal of Physics: Conference Series 2083 (2021) 032030, doi:10.1088/1742-6596/2083/3/032030
7. Ghimire, D., Jeong, S., Lee, J. *et al.* (2017). Facial expression recognition based on local region specific features and support vector machines. *Multimed Tools Appl* 76, pp. 7803–7821.
8. Gudi, A., Tasli, H.E., Den Uyl, T.M., Maroulis, A. (2015). Deep learning based facs action unit occurrence and intensity estimation. 11th IEEE international conference and workshops on automatic face and gesture recognition.
9. Kim, D. H., Baddar, W. J., Jang, J., Ro, Y. M. (2019). Multi-Objective Based Spatio-Temporal Feature Representation Learning Robust to Expression Intensity Variations for Facial Expression Recognition. *IEEE Transactions on Affective Computing*, *10*(2), pp. 223–236.
10. Ko, B.C. (2018). A Brief Review of Facial Emotion Recognition Based on Visual Information. *Sensors* 2018, *18*, 401. https://doi.org/10.3390/s18020401
11. Levi, G. & Hassner, T. (2015). Emotion Recognition in the Wild via Convolutional Neural Networks and Mapped Binary Patterns. *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*.
12. Li, S. & Deng, W. (2020). Deep Facial Expression Recognition: A Survey, *IEEE Transactions on Affective Computing*, doi: 10.1109/TAFFC.2020.2981446, 2020
13. Lowe, D. G., (1999). Object recognition from local scale-invariant features," Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150-1157.
14. Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J. (1998). Coding facial expressions with Gabor wavelets, Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200-205.
15. Minaee, S., Minaei, M., Abdolrashidi, A. (2021). Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network. *Sensors 2021*, *21*, 3046. https://doi.org/10.3390/s21093046
16. Ojala, T., Pietikainen, M., Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 7, pp. 971-987, July 2002, doi: 10.1109/TPAMI.2002.1017623.
17. Pons, G. & Masip, D. (2018). Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition, doi:10.48550/ARXIV.1802.06664,
18. Sariyanidi, E., Gunes, H., Cavallaro, A. (2015). Automatic Analysis of Facial Affect: A Survey of Registration, Representation, and Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(6), 1113–1133.
19. Usman, M., Latif, S., Qadir, J. (2017). Using deep autoencoders for facial expression recognition, *13th International Conference on Emerging Technologies (ICET)*, pp. 1-6.
20. Viola, P., Jones, M.J. (2004). Robust Real Time Face Detection, International Journal of Computer Vision 57, pp. 137-154.
21. Zafeiriou, S., Zhang, C., Zhang, Z. (2015). A survey on face detection in the wild: Past, present and future, Computer Vision and Image Understanding, Vol. 138, pp 1-24.
22. Zeng, N., Zhang, H., Song, B., Liu, W., Li, Y., Abdullah, D. (2018). Facial Expression Recognition via Learning Deep Sparse Autoencoders, Neurocomputing, vol. 273, pp. 643-649.
23. Zhang, C., & Zhang, Z. (2014). Improving multiview face detection with multi-task deep convolutional neural networks. *IEEE Winter Conference on Applications of Computer Vision*, 1036-1041.
24. Zhang, T., Zheng, W., Cui, Z., Zong, Y., Yan, J., Yan, K. (2016). A Deep Neural Network Driven Feature Learning Method for Multi-view Facial Expression Recognition. IEEE Transactions on Multimedia. 18. 1-1.

25. Zhu, X. & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild, 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2879-2886, doi: 10.1109/CVPR.2012.6248014.

Appendix A. Video file analysis

GUT-C01-S01-20210618

Code	FIT_FAILED	FIND_FAILED	DETECTED
GUT-C01-S01-20210618	6.42%	8.38%	85.20%

GUT-C01-S02-20210621

Code	FIT_FAILED	FIND_FAILED	DETECTED
GUT-C01-S02-20210621	21.21%	17.62%	61.17%

GUT-C01-S03-20210628

Code	FIT_FAILED	FIND_FAILED	DETECTED
GUT-C01-S03-20210628	30.32%	13.32%	56.36%

Notes:

- At the end child often looked down on the controller

GUT-C02-S01-20210618

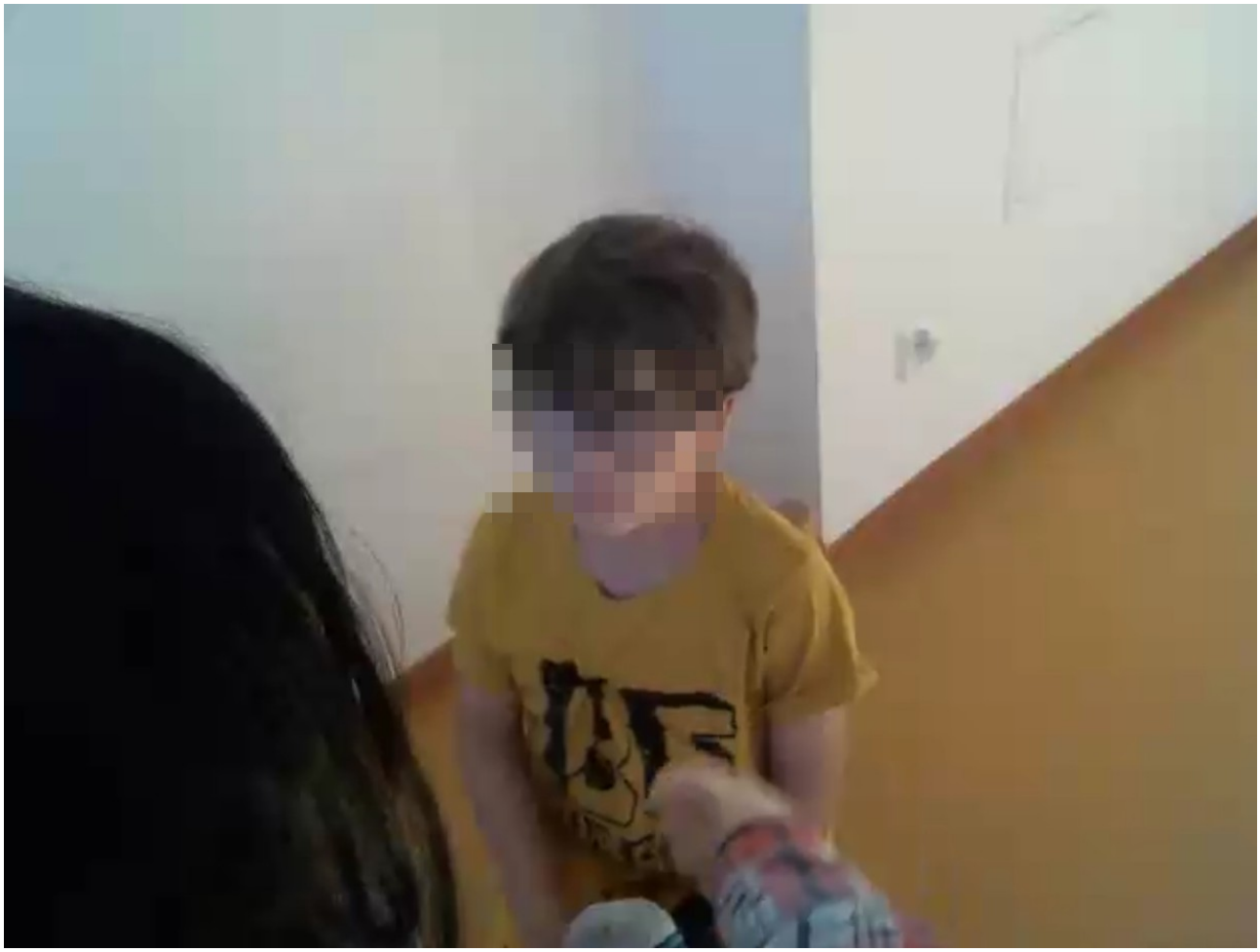
Code	FIT_FAILED	FIND_FAILED	DETECTED
GUT-C02-S01-20210618	22.52%	75.13%	2.35%

Notes:

- The child often lowers his head while looking at his lap.
- The child has thick glasses and the frame of the glasses often covers the eyes. A long fringe covers the forehead.
- At the end of the video, the child approached the robot.

GUT-C02-S02-20210621

Code	FIT_FAILED	FIND_FAILED	DETECTED
GUT-C02-S02-20210621	10.79%	89.05%	0.16%



Notes:

- The child has thick glasses and the frame of the glasses often covers the eyes. A long fringe covers the forehead.
- For almost the entire film, the child does not look at the robot but to the side or down

GUT-C02-S03-20210628

Code	FIT_FAILED	FIND_FAILED	DETECTED
GUT-C02-S03-20210628	20.28%	71.00%	8.71%

Notes:

- The child has thick glasses and the frame of the glasses often covers the eyes. A long fringe covers the forehead.
- The child looks down and to the side, but less than in previous sessions
- At the end of the video, no face is visible because the boy was looking down at the controller.

GUT-C03-S01-20210618

Code	FIT_FAILED	FIND_FAILED	DETECTED
GUT-C03-S01-20210618	6.46%	9.05%	84.49%

Notes:

- Child often partially covers face with hand.

GUT-C03-S02-20210621

Code	FIT_FAILED	FIND_FAILED	DETECTED
GUT-C03-S02-20210621	20.37%	41.49%	38.14%

Notes:

- The child often plays with his hair and touches his face thus hiding his face
- Towards the end the camera could not see his face because the boy was looking down at the controller

MAAP-C01-S01-20200907

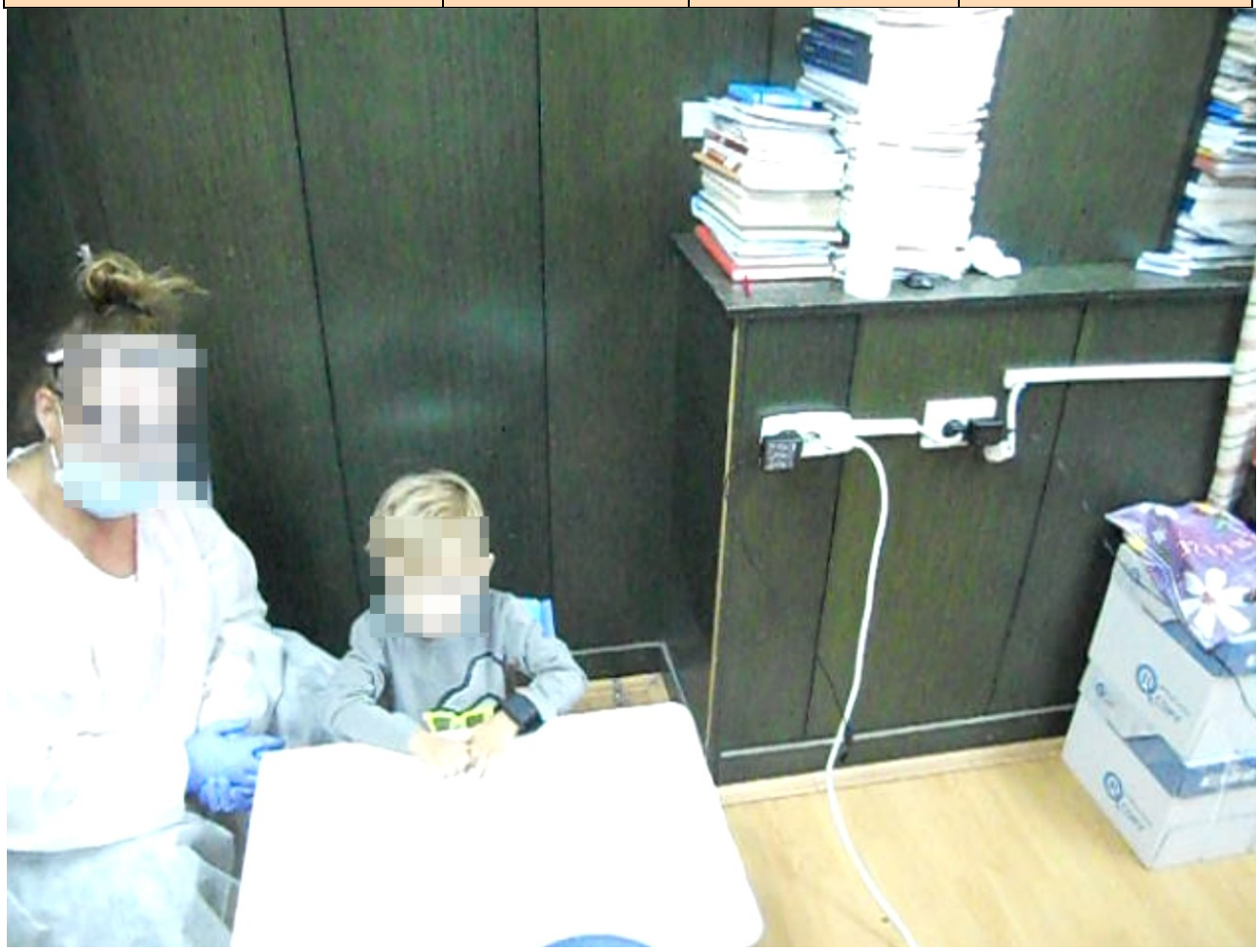
Code	FIT_FAILED	FIND_FAILED	DETECTED
MAAP-C01-S01-20200907	36.53%	63.32%	0.15%

Notes:

- Low resolution
- In addition to the child, another person can also be seen
- Very bright light, overexposed video
- Wide crop - long distance from the child's face

MAAP-C01-S02-20200914

Code	FIT_FAILED	FIND_FAILED	DETECTED
MAAP-C01-S02-20200914	82.23%	7.34%	10.43%



Notes:

- Low resolution
- Very bright light, overexposed video
- In addition to the child, another person can also be seen
- Camera view from above, too high angle

MAAP-C01-S03-20200921

Code	FIT_FAILED	FIND_FAILED	DETECTED
MAAP-C01-S03-20200921	56.49%	23.72%	19.80%

Notes:

- Low resolution
- Very bright light, overexposed video
- In addition to the child, another person can also be seen
- Camera view from above, too high angle

MAAP-C01-S04-20201005

Code	FIT_FAILED	FIND_FAILED	DETECTED
MAAP-C01-S04-20201005	60.05%	22.11%	17.84%

Notes:

- Low resolution
- Very bright light, overexposed video
- In addition to the child, another person can also be seen
- Camera view from above, too high angle

MAAP-C02-S01-20200908

Code	FIT_FAILED	FIND_FAILED	DETECTED
MAAP-C02-S01-20200908	24.97%	74.92%	0.11%

Notes:

- Low resolution
- In addition to the child, another persons can also be seen
- Very bright light, overexposed video
- Wide crop - long distance from the child's face

MAAP-C02-S02-20200915

Code	FIT_FAILED	FIND_FAILED	DETECTED
MAAP-C02-S02-20200915	59.42%	33.88%	6.70%

Notes:

- Low resolution
- Very bright light, overexposed video
- In addition to the child, another person can also be seen
- Camera view from above, too high angle

MAAP-C02-S03-20200917

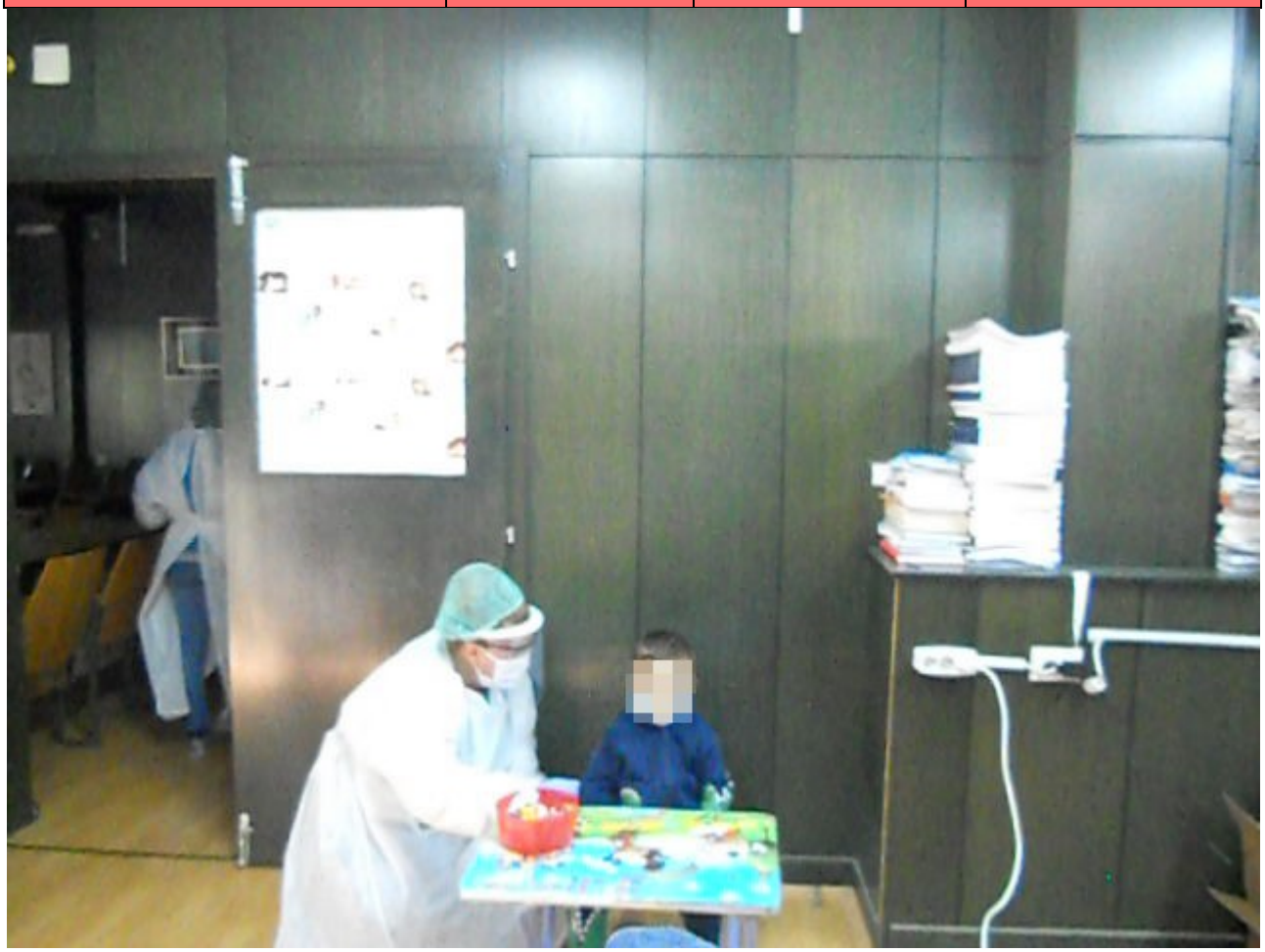
Code	FIT_FAILED	FIND_FAILED	DETECTED
MAAP-C02-S03-20200917	37.26%	59.39%	3.35%

Notes:

- Low resolution
- Very bright light, overexposed video
- In addition to the child, another person can also be seen
- Camera view from above, too high angle

MAAP-C03-S01-20200907

Code	FIT_FAILED	FIND_FAILED	DETECTED
MAAP-C03-S01-20200907	28.64%	71.27%	0.09%



Notes:

- Low resolution
- In addition to the child, another persons can also be seen
- Very bright light, overexposed video
- Wide crop - long distance from the child's face

MAAP-C03-S02-20200914

Code	FIT_FAILED	FIND_FAILED	DETECTED
MAAP-C03-S02-20200914	20.39%	8.15%	71.46%

Notes:

- Low resolution
- In addition to the child, another person can also be seen
- Camera view from above, too high angle

MAAP-C03-S03-20200915

Code	FIT_FAILED	FIND_FAILED	DETECTED
MAAP-C03-S03-20200915	23.02%	1.97%	75.00%

Notes:

- High quality camera
- Camera view from above, too high angle

MAAP-C03-S04-20200921

Code	FIT_FAILED	FIND_FAILED	DETECTED
MAAP-C03-S04-20200921	33.73%	6.94%	59.33%

Notes:

- Low resolution
- In addition to the child, another person can also be seen
- Camera view from above, too high angle

MAAP-C03-S05-20200922

Code	FIT_FAILED	FIND_FAILED	DETECTED
MAAP-C03-S05-20200922	36.97%	8.81%	54.22%

Notes:

- Low resolution
- In addition to the child, another person can also be seen
- Camera view from above, too high angle

MAAP-C03-S06-20200928

Code	FIT_FAILED	FIND_FAILED	DETECTED
MAAP-C03-S06-20200928	33.84%	12.52%	53.64%

Notes:

- Low resolution
- In addition to the child, another person can also be seen
- Camera view from above, too high angle

MAAP-C03-S07-20200929

Code	FIT_FAILED	FIND_FAILED	DETECTED
MAAP-C03-S07-20200929	21.79%	12.25%	65.97%

Notes:

- Low resolution
- In addition to the child, another person can also be seen
- Camera view from above, too high angle

MAAP-C03-S08-20201005

Code	FIT_FAILED	FIND_FAILED	DETECTED
MAAP-C03-S08-20201005	58.28%	39.47%	2.24%

Notes:

- Very low resolution
- In addition to the child, another person can also be seen
- Camera view from above, too high angle

MAAP-C03-S09-20201006

Code	FIT_FAILED	FIND_FAILED	DETECTED
MAAP-C03-S09-20201006	56.08%	37.77%	6.15%

Notes:

- Very low resolution
- In addition to the child, another person can also be seen
- Camera view from above, too high angle

MAAP-C03-S10-20201007

Code	FIT_FAILED	FIND_FAILED	DETECTED
MAAP-C03-S10-20201007	55.64%	41.59%	2.77%

Notes:

- Very low resolution
- In addition to the child, another person can also be seen
- Camera view from above, too high angle

UH-C01-S01-20210624

Code	FIT_FAILED	FIND_FAILED	DETECTED
UH-C01-S01-20210624	44.43%	51.08%	4.49%

Notes:

- Very bright light, overexposed video
- Camera angled too much both vertically and horizontally

UH-C01-S02-20210625

Code	FIT_FAILED	FIND_FAILED	DETECTED
UH-C01-S02-20210625	46.83%	33.42%	19.74%

Notes:

- Camera angled too much both vertically and horizontally
- Kaspar's hat sometimes obscures the child's face

UH-C01-S03-20210629

Code	FIT_FAILED	FIND_FAILED	DETECTED
UH-C01-S03-20210629	33.53%	51.02%	15.45%

Notes:

- Camera angled too much both vertically and horizontally
- The child's face goes out of frame

UH-C02-S01-20210624

Code	FIT_FAILED	FIND_FAILED	DETECTED
UH-C02-S01-20210624	18.81%	21.73%	59.46%

UH-C03-S01-20210624

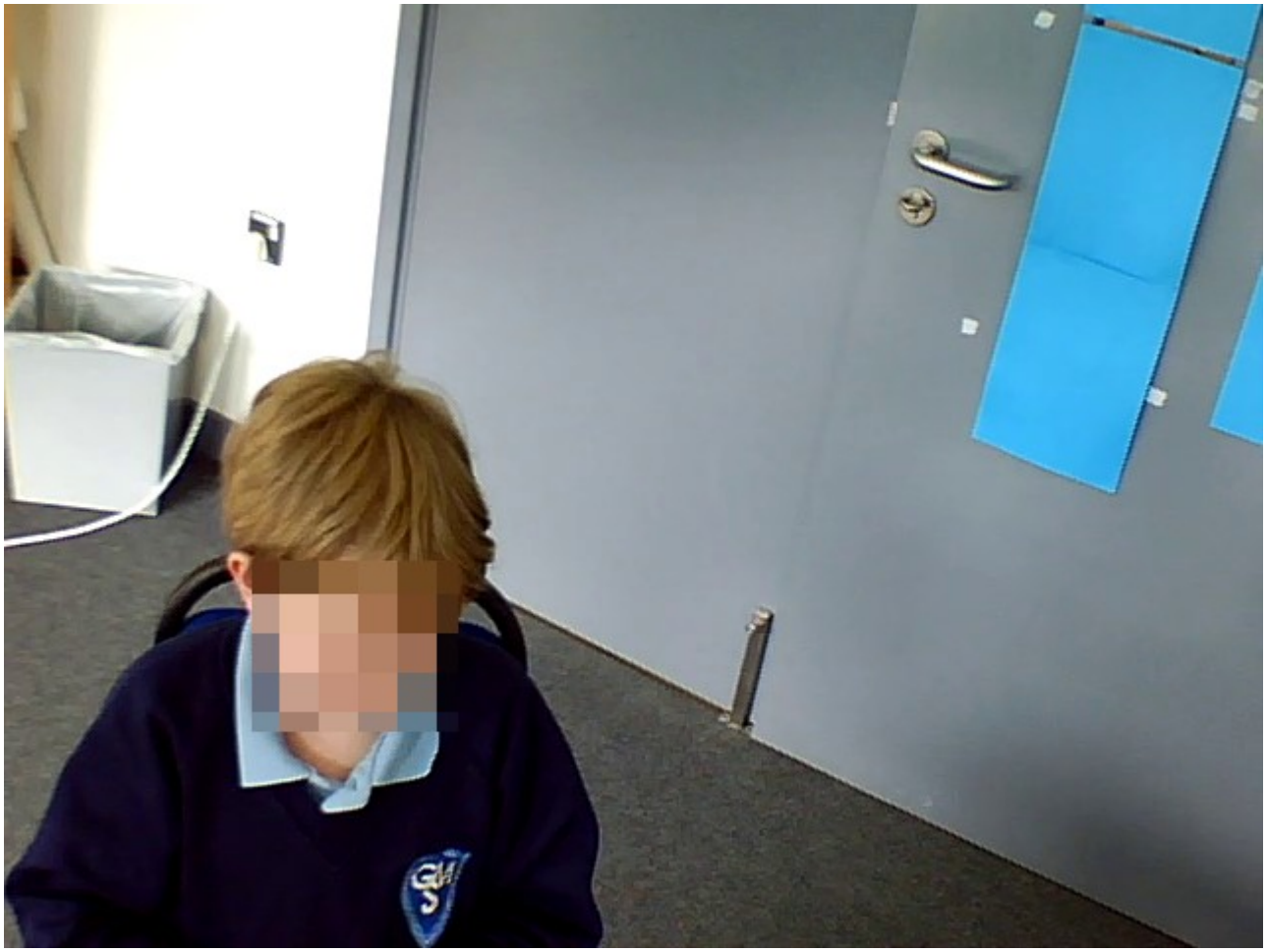
Code	FIT_FAILED	FIND_FAILED	DETECTED
UH-C03-S01-20210624	25.13%	74.50%	0.37%

Notes:

- Very bright light, overexposed video
- Camera angled too much both vertically and horizontally
- A child's long fringe

UH-C03-S02-20210629

Code	FIT_FAILED	FIND_FAILED	DETECTED
UH-C03-S02-20210629	12.21%	79.35%	8.44%



Notes:

- Camera angled too much both vertically and horizontally
- A child's long fringe

UH-C04-S01-20210624

Code	FIT_FAILED	FIND_FAILED	DETECTED
UH-C04-S01-20210624	8.00%	90.94%	1.06%

Notes:

- Very bright light, overexposed video
- Camera angled too much both vertically
- A child's long fringe
- She often covers his face with his hands

UH-C04-S02-20210625

Code	FIT_FAILED	FIND_FAILED	DETECTED
UH-C04-S02-20210625	33.09%	52.35%	14.56%

Notes:

- Camera angled too much both vertically
- A child's long fringe

UH-C04-S03-20210629

Code	FIT_FAILED	FIND_FAILED	DETECTED
UH-C04-S03-20210629	14.75%	84.23%	1.01%

Notes:

- Very bright light, overexposed video
- Camera angled too much both vertically
- A child's long fringe

UH-C05-S01-20210624

Code	FIT_FAILED	FIND_FAILED	DETECTED
UH-C05-S01-20210624	4.01%	9.38%	86.60%

UH-C05-S02-20210625

Code	FIT_FAILED	FIND_FAILED	DETECTED
UH-C05-S02-20210625	0.00%	100.00%	0.00%

Notes:

- The clip is only one second long.

UH-C05-S03-20210629

Code	FIT_FAILED	FIND_FAILED	DETECTED
UH-C05-S03-20210629	31.52%	16.85%	51.63%

UH-C06-S01-20210624

Code	FIT_FAILED	FIND_FAILED	DETECTED
UH-C06-S01-20210624	20.95%	73.65%	5.40%

Notes:

- Glasses slipped down the nose
- Kaspar's hat often covers the child's face

UH-C06-S02-20210629

Code	FIT_FAILED	FIND_FAILED	DETECTED
UH-C06-S02-20210629	60.99%	16.48%	22.53%

Notes:

- Glasses slipped down the nose

- **ITU-C01-S01-20210605**

Code	FIT_FAILED	FIND_FAILED	DETECTED
ITU-C01-S01-20210605	13,96%	81,89%	4,15%

Notes:

- Face mask
- You can see the therapist's face in the recording
- Camera angle too high

ITU-C01-S02-20210702

Code	FIT_FAILED	FIND_FAILED	DETECTED
ITU-C01-S02-20210702	10,68%	85,89%	3,43%

Notes:

- You can see the therapist's face in the recording
- Camera angle too high

ITU-C02-S01-20210619

Code	FIT_FAILED	FIND_FAILED	DETECTED
ITU-C02-S01-20210619	13,28%	85,42%	1,30%

Notes:

- Kaspar obscures the child's face
- You can see the therapist's face in the recording
- Camera angle too high

ITU-C03-S01-20210619

Code	FIT_FAILED	FIND_FAILED	DETECTED
ITU-C03-S01-20210619	25,41%	22,78%	51,81%

Notes:

- The child's face sometimes goes out of frame
- You can see the therapist's face in the recording
- Camera angle too high

ITU-C04-S01-20210619

Code	FIT_FAILED	FIND_FAILED	DETECTED
ITU-C04-S01-20210619	15,08%	61,41%	23,51%

Notes:

- The child's face sometimes goes out of frame
- Camera angle too high

ITU-C05-S01-20210702

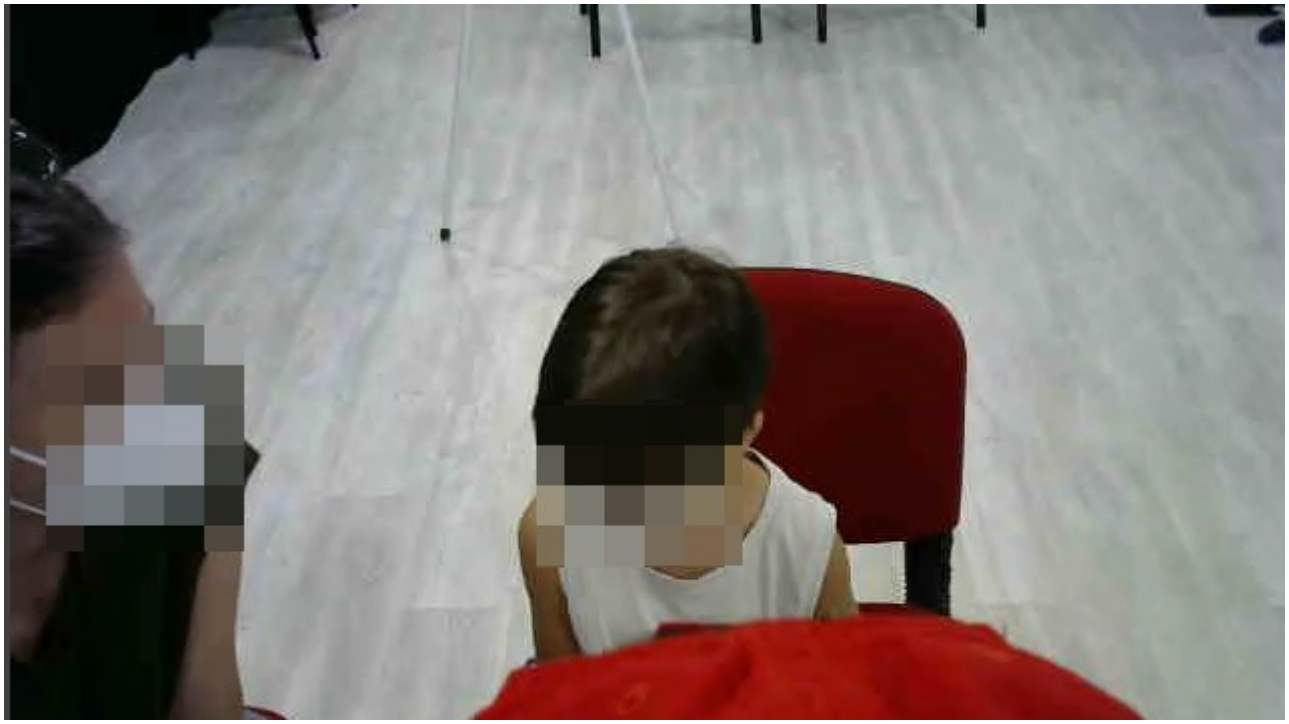
Code	FIT_FAILED	FIND_FAILED	DETECTED
ITU-C05-S01-20210702	12,37%	57,38%	30,25%

Notes:

- Face mask - halfway through the film slipped off
- You can see the therapist's face in the recording
- Camera angle too high

ITU-C06-S01-20210810

Code	FIT_FAILED	FIND_FAILED	DETECTED
ITU-C06-S01-20210810	2,15%	97,82%	0,03%



Notes:

- Face mask
- You can see the therapist's face in the recording
- Camera angle too high
- A child's long fringe

ITU-C07-S01-20210810

Code	FIT_FAILED	FIND_FAILED	DETECTED
ITU-C07-S01-20210810	14,33%	23,23%	62,44%

Notes:

- Face mask - halfway through the film slipped off
- You can see the therapist's face in the recording
- Camera angle too high

ITU-C08-S01-20210810

Code	FIT_FAILED	FIND_FAILED	DETECTED
ITU-C08-S01-20210810	3,44%	95,52%	1,04%

Notes:

- Face mask
- You can see the therapist's face in the recording
- Camera angle too high
- Lush haircut

ITU-C09-S01-20210810

Code	FIT_FAILED	FIND_FAILED	DETECTED
ITU-C09-S01-20210810	13,12%	81,39%	5,49%

Notes:

- Face mask - halfway through the film slipped off
- You can see the therapist's face in the recording
- Camera angle too high

- **ITU-C10-S01-20210810**

Code	FIT_FAILED	FIND_FAILED	DETECTED
ITU-C10-S01-20210810	5,89%	66,21%	27,90%

Notes:

- Face mask - halfway through the film slightly slipped off
- You can see the therapist's face in the recording
- Camera angle too high
- Kaspar obscures the child's face

ITU-C12-S01-20210810

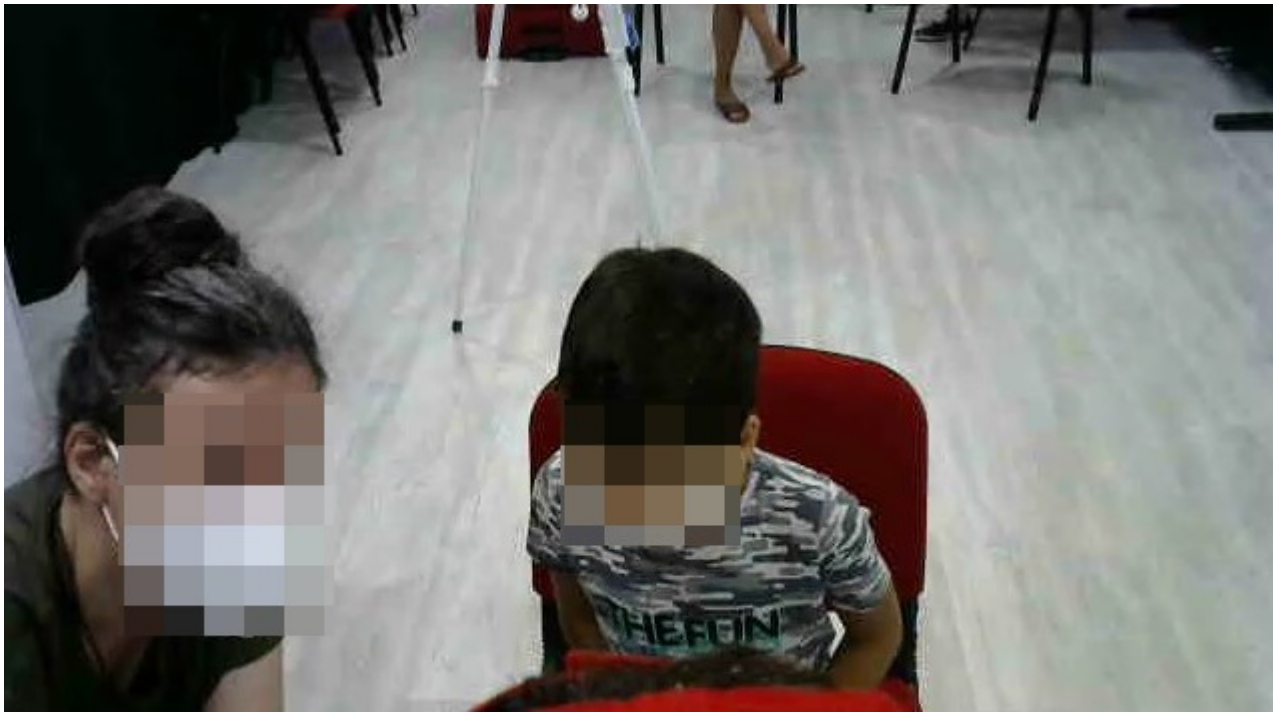
Code	FIT_FAILED	FIND_FAILED	DETECTED
ITU-C12-S01-20210810	14,61%	33,30%	52,09%

Notes:

- You can see the therapist's face in the recording
- Camera angle too high

ITU-C13-S01-20210810

Code	FIT_FAILED	FIND_FAILED	DETECTED
ITU-C13-S01-20210810	21,01%	78,51%	0,48%



Notes:

- Face mask - halfway through the film slightly slipped off
- You can see the therapist's face in the recording
- Camera angle too high
- Kaspar obscures the child's face
- **d**