Medical Face Masks and Emotion Recognition from the Body: Insights from a Deep Learning Perspective

Nikolaos Kegkeroglou
National Technical University of Athens, School of ECE
15773 Athens, Greece
nkegkeroglou@gmail.com

Panagiotis P. Filntisis
Athena Research Center
Institute of Robotics
15125 Maroussi, Greece
filby@central.ntua.gr

Petros Maragos
National Technical University of Athens, School of ECE
15773 Athens, Greece
Athena Research Center
Institute of Robotics
15125 Maroussi, Greece
maragos@cs.ntua.gr

ABSTRACT

The COVID-19 pandemic has undoubtedly changed the standards and affected all aspects of our lives, especially social communication. It has forced people to extensively wear medical face masks, in order to prevent transmission. This face occlusion can strongly irritate emotional reading from the face and urges us to incorporate the whole body as an emotional cue. In this paper, we conduct insightful studies about the effect of face occlusion on emotion recognition performance, and showcase the superiority of full body input over the plain masked face. We utilize a deep learning model based on the Temporal Segment Network framework, and aspire to fully overcome the face mask consequences. Although facial and bodily features can be learned from a single input, this may lead to irrelevant information confusion. By processing those features separately and fusing their prediction scores, we are more effectively taking advantage of both modalities. This framework also naturally supports temporal modeling, by mingling information among neighboring frames. In combination, these techniques form an effective system capable of tackling emotion recognition difficulties, caused by safety protocols applied in crucial areas.

CCS CONCEPTS

• **Computing methodologies → Computer vision problems**.

KEYWORDS

Body Expression, Visual Emotion Recognition, Child-Robot Interaction, COVID-19, Deep Learning

ACM Reference Format:
Nikolaos Kegkeroglou, Panagiotis P. Filntisis, and Petros Maragos. 2023. Medical Face Masks and Emotion Recognition from the Body: Insights from a Deep Learning Perspective. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '23), July 5–7, 2023, Corfu, Greece.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3594806.3594829

1 INTRODUCTION

The possible applications of an interface capable of assessing human emotional states are numerous. Humans generally treat computer agents as they might treat other people [32]. Robots and systems that are able to recognize, interpret and process human affect [5], are arguably well suited to this, making the interaction more effective and pleasant. They find fertile ground in the area of computer-assisted education, as learning is the quintessential emotional experience. A learning episode might begin with curiosity and fascination. But as its difficulty increases, one may experience confusion, frustration or anxiety, and thus, may abandon learning [29]. A tutoring agent, who is able to estimate the learner's affective state, can respond appropriately and give encouraging suggestions. Existing work has shown that robot tutors enhance learning, by personalizing their motivational strategies to the student's emotional behavior [13] [20]. Another crucial area is health care, as mental health disorders, like depression and psychosis, are on the rise across the world. Emotion recognition systems can be an effective strategy for preventing and monitoring such disorders [38].

While works based on facial expressions abound in the area, recognizing affect from the body remains a less explored topic. A study in neurobiology has shown that body movement and posture contain useful features for recognizing human affect [8]. In other experiments, it was shown that facial and bodily expressions work complementary for visual perception of emotion, and in some cases humans perceive bodily expressed emotional information as more diagnostic than facial [2]. Furthermore, the visibility of facial cues is not guaranteed. Bodily expression recognition becomes crucial when facial features are occluded. Medical face masks, which are extensively used nowadays due to the COVID-19 pandemic [6], are the epitome of face occlusion. Because bodies are more expressive than faces in those situations, social information can be detected from the body instead.

Although there has been a considerable amount of research on automatic emotion recognition in adults, the topic regarding children has been understudied. Children go through a critical development process and applications involving them require special attention [31]. They also tend to fidget and move around more than adults, leading to more self-occlusions and non-frontal head poses [4]. This becomes even more challenging, considering the current health and safety protocols that demand the use of face masks.

PETRA '23, July 5–7, 2023, Corfu, Greece

ACM ISBN 979-8-4007-0069-9/23/07…$15.00
https://doi.org/10.1145/3594806.3594829

Proc. Int'l Conference on PErvasive Technologies Related to Assistive Environments (**PETRA 2023**), Corfu, Greece, July 2023.

Robots can no longer rely only on facial expressions to recognize emotion, but also have to take into account body expressions that can stay visible and detectable, even when the face is unobservable. Children's behavior and natural characteristics differ from adults, so perception systems need to be specifically trained, to be able to tackle Child-Robot Interaction (CRI) problems.

The rest of the paper is organized as follows: Section 2 discusses related work on emotion recognition mainly from the body. Section 3 describes the adopted deep learning-based visual emotion recognition model in detail, as well as the tools and methods used for the experiments. Section 4 presents the experimental results, and lastly Section 5 provides our conclusions.

2 RELATED WORK

In recent years, deep learning methods have been very popular due to the massive amounts of digital data in combination with powerful processing hardware. Deep extracted features have yielded excellent results and on most cases outperformed non-deep state-of-the-art methods for the emotion recognition task. When processing a video with emotional expressions, an essential component is capturing temporal information to complement the prediction from still images. Two-stream Convolutional Neural Network (CNN) architectures use multi-frame optical flow to handle complex actions like emotional expressions [34].

The most common modality used by the research community for identifying emotion is facial expressions [21]. Some works have proposed an audiovisual approach [12], where the system takes speech as an additional input to the face, in order to tackle occlusions and increase robustness. In [26], they utilize 3D CNNs to extract spatio-temporal features both from face videos and audio signals, and deep belief nets [17] for emotion recognition. However, the COVID-19 pandemic has fostered a pervasive use of medical face masks all around the world, making a serious impact on social communication. Several studies investigated how the presence of a face mask affects emotion recognition accuracy and revealed that it diminishes the people's ability to accurately categorize a facial expression [6] [14]. On top of that, the mask impairs re-identification of the same face by people [24], which suggests a need for mask-specific model training. In [35], they also explored how masks influence the perceived emotional profile of facial expressions. It was shown, that it not only led to a decrease in perceived intensity of the intended emotions, but it also resulted in an overall increase in the perceived intensity of non-intended emotions. In [28], even super-recognizers, people who are highly skilled and superior in recognition tasks, were impaired by the face occlusion caused by the face mask. This negative effect in emotional reading is not limited to adults, as it also concerns interaction with children [7].

Motivated by all the above, we move towards incorporating bodily expressed information as a major cue in the emotion recognition task. An early work [15] combined handcrafted face and body features at feature and decision-level for emotion classification. In [10], a hierarchical multi-label annotation method was proposed, which fused body skeleton with facial expressions for automatic recognition of emotion of children during CRI scenarios. In [23], they experimented with two bodily expression pipelines, one of which implemented a two-stream-based CNN. The other one relied solely on the human skeleton and utilized a spatial-temporal Graph Convolutional Network (GCN) [37], which constructs a graph from the human body landmarks with their natural spatial connectivity, as well as temporally neighboring landmarks.

Along with body, context has been an additional modality involved in the task of emotion recognition. In [11], RGB and flow body streams were accompanied with a context RGB stream and a visual-semantic embedding loss based on word embedding representations. In [18], they proposed a network structure composed of a GCN processing skeleton landmarks, and two 3D CNNs for RGB body and context input. A network ensemble, including streams that processed the body in RGB, flow and skeleton form was proposed in [30]. This variety of bodily expressed cues have also been involved in CRI emotion recognition systems [9] [25].

Our work focuses on the effect of the face occlusion on emotion recognition performance. We adopt a proven related work model and process only RGB input, despite the diversity of body cues that can be conveyed. We sense that this approach suits best to our purpose, regarding the medical face mask effect study.

3 VISUAL EMOTION RECOGNITION MODEL

In this chapter, we present the model, that will be used to tackle the visual emotion recognition task. We discuss its structure and benefits and also address the occuring challenges, which are taken into account in the model's various configurations. Furthermore, we describe the tools utilized to conduct the upcoming experiments and some techniques to enhance model performance.

3.1 Feature Capturing

Complex actions, like emotional expressions, comprise multiple stages spanning over a period of time and it would be quite a loss failing to utilize them. On the other hand, each expressed emotion is not present throughout a whole input video. These facts, indicate that we are in need of effective general feature capturing. While the plain CNN architecture considers the whole input sequence, as well as each frame in the video separately, the Temporal Segment Network (TSN) framework [36] operates on a sequence of short snippets sparsely sampled from the entire video. Each snippet in this sequence will produce its own preliminary prediction of the emotion classes and then, a consensus among the snippets will be derived as the video-level prediction. Therefore, it allows the network to access several parts of the video, but also tackles the inability of the former to model long-range temporal structure, thus, being more likely to observe the corresponding expression.

3.2 Method

The overall architecture of our model is shown in Fig. 1. Formally, given a video V, we divide it into K non-overlapping segments $\{S_1, S_2, ..., S_K\}$, to access several parts of the video, and transform them into a sequence of snippets $\{T_1, T_2, ..., T_K\}$. Each snippet T_k is produced, by randomly sampling 3 consecutive frames from its corresponding segment S_k, to tackle frame redundancy. Finally, a segmental consensus function $\mathcal{H}$ is applied on the snippet-level predictions produced by the backbone, to obtain the final scores S:

$$S = \mathrm{TSN}(T_1, T_2, ..., T_K) = \mathcal{H}(\mathcal{F}(T_1; \mathbf{W}), \mathcal{F}(T_2; \mathbf{W}), ..., \mathcal{F}(T_K; \mathbf{W}))$$

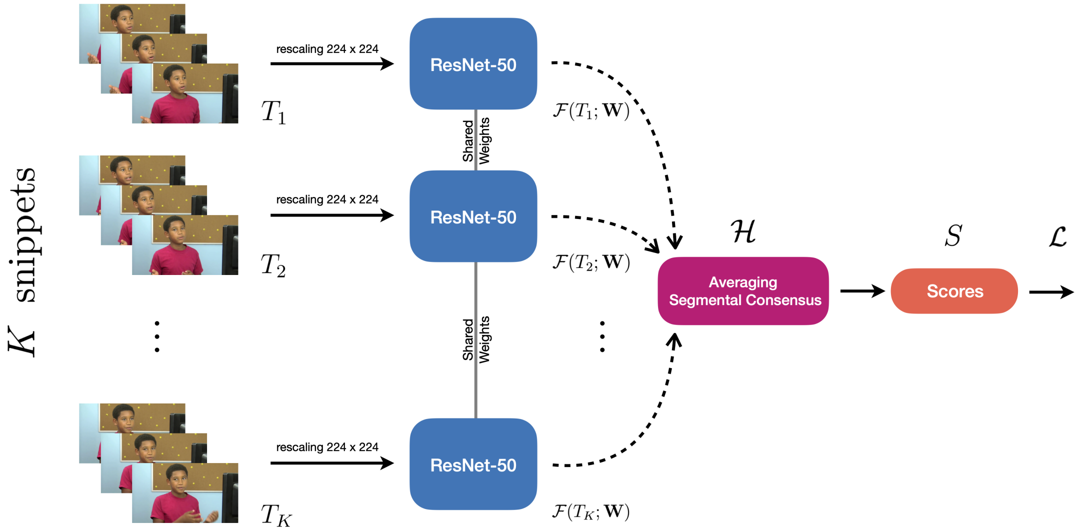


Figure 1: TSN-Based Model Architecture

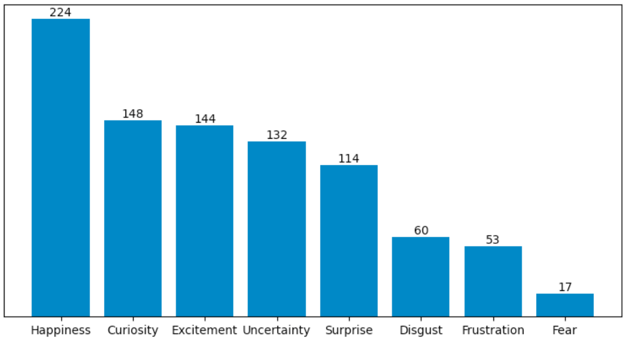


Figure 2: EmoReact Training Set Imbalance

Here $\mathcal{F}(T_k; \mathbf{W})$ denotes the function representing the application of a CNN with parameters $\mathbf{W}$ on the snippet T_k. The CNN is equipped with a ResNet-50 backbone architecture [16]. The consensus function $\mathcal{H}$ we use is average pooling and the obtained video-level scores S are fed to a loss function $\mathcal{L}$ to perform the training step. This framework offers several benefits to emotion recognition. Compared to processing the entire video, the sampling process ignores redundant information in consecutive video frames, helping avoid overfitting, and offers a type of data augmentation, valuable for children emotion databases of small size.

3.3 Database

We perform our experiments on the EmoReact dataset [27], which contains 1102 videos of 63 children, aged between 4 and 14, expressing emotions while discussing different topics, collected from the YouTube channel React. Each video is annotated with one or more emotions, from a total of 8 emotion labels: Curiosity, Uncertainty, Excitement, Happiness, Surprise, Disgust, Fear, and Frustration. Therefore, we are dealing with a CRI binary multi-label classification problem. In Fig. 2, we show the imbalance of EmoReact's training set, which means it includes an unequal number of videos for each emotion label, and is something that we must address in our upcoming model configuration choices. We can also argue that some emotions (Fear, Frustration, Disgust) are expressed in a relatively low number of samples, which results in possible lack of diversity and less ease to generalize well across unseen individuals, introducing an extra degree of difficulty to our problem.

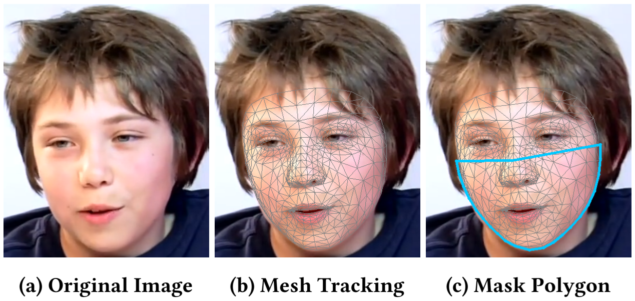

(a) Original Image (b) Mesh Tracking (c) Mask Polygon

Figure 3: Mask Application Steps

3.4 Medical Face Mask Effect Study

The COVID-19 pandemic has forced people to extensively wear medical face masks, in order to prevent transmission. Motivated by this fact, we want to conduct an experimental study about the effect of medical face masks on emotion recognition, by applying a relevant mask on the EmoReact children's faces, as an attempt to simulate the face occlusion consequence.

3.4.1 Mask Application. To apply the mask, we detect the facial surface geometry using Google's MediaPipe Face Mesh [19], an end-to-end CNN-based model for inferring an approximate 3D mesh representation of a human face from a single image. It uses a dense mesh model of 468 vertices and is well-suited for face-based augmented reality effects. We track the 2D coordinates of the right

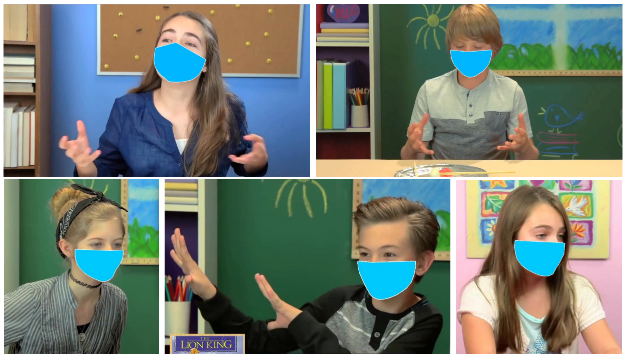


Figure 4: EmoReact Masked Samples

and left jawline vertices, starting from just below the eyes until the chin, and one extra vertex for the nose, in order to form a polygon that is finally filled to represent the mask (Fig. 3). The jawlines for the mask are created by tracking the edge x-axis vertices and accordingly selecting among several jawline candidates, that we manually created for this particular face mesh model [1]. In Fig. 4, we display several samples of EmoReact after the application of the mask and showcase our tool's robustness to face orientation.

3.4.2 Body Detection. In order to incorporate bodily expressions, we need a way to detect the human body. Google's MediaPipe also provides human body and hand skeleton tracking tools [3] [39]. We combine keypoints tracked by both tools and create a bounding box with the edge points, expanded by a factor of 10% at each respective dimension, which is then cropped as the input image. This process is demonstrated in Fig. 5, where most background noise is removed and full body information dominates the cropped image.

3.5 Modality Fusion

We are looking to take advantage of the face and body information separately, by fusing the individual modality prediction scores with a late fusion scheme. The full body crop includes the masked face, and processing it as a single RGB input image can lead to irrelevant information confusion. The proposed method is to separate the face and body features, in order to avoid the aforementioned issue. The core model remains as is, but now processes the face crop, and the plain body crop with the corresponding face area blacked out, in two separate forward passes (Fig. 6). After producing the scores S_f and S_b from face and plain body respectively, we use a late fusion scheme to obtain the final scores S. Finally, the overall loss $\mathcal{L}$ is simply the summation of the individual modality losses: $\mathcal{L} = L_f + L_b$.

3.6 Temporal Modeling

The current TSN-based model processes only one of the N consecutive frames of each snippet, being heavily based on spatial structure. This architecture naturally supports temporal modeling, by mingling information among neighboring snippet frames with the Temporal Shift Module (TSM) [22]. TSM can be inserted into CNNs, to exploit temporality at zero computation and parameters. It shifts part of the channels of the input frames and the latent representations of each snippet along the temporal dimension, both forward and backward, thus facilitate information exchange among neighboring frames. Because information contained in the channels is no longer accessible for the current frame, the spatial modeling ability of the backbone can be harmed. To address this problem, the module is placed inside the residual branches of the ResNet, so the information in the original activation is still accessible after temporal shift, through the identity mappings.

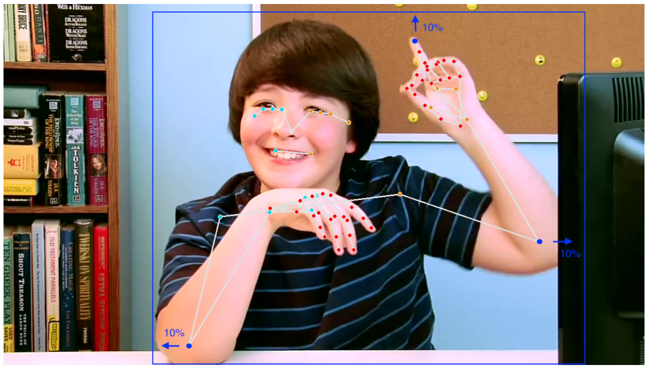


Figure 5: Body Detection

3.7 Model & Training Configurations

The model is pretrained on AffectNet, the largest facial expression dataset. We obtain the weights of the network as provided by the PyTorch framework, achieving 59.47% accuracy on the validation set. Before feeding the input to the network, we rescale sampled RGB images from full resolution to 224 × 224. We train our models for 60 epochs, with stochastic gradient descent with momentum 0.9 and a batch size of 8, L2 regularization with weight decay 5e-4, and start with a learning rate of 1e-2, which is then reduced by a factor of 10 at 20 and 40 epoch milestones[2]. Since our task is binary multi-label classification, our predictions are fed to a binary cross-entropy (BCE) loss function, after suppressing the scores S to [0,1] with a sigmoid function. BCE depends on the the label-specific error, thus it penalizes label predictions independently. Following prior work, the only evaluation metric that has been shown to be robust to imbalanced datasets is the Area Under the Curve of Receiver Operating Characteristic (ROC AUC). The scores S, of size 8 per sample, are averaged to obtain a single overall performance metric. Instead of giving equal weight to each class, which will over-emphasize on the typically low performance on an infrequent class, we compute the unbalanced average, so every sample-class pair contributes equally to the overall metric. For evaluation, we select the epoch with the best validation ROC AUC and apply the corresponding network on the test set, to finally report the best overall performance achieved.

[1] The code for the mask application tool is publicly available at: https://github.com/nkegke/medical-face-mask-applier

[2] The code for the model and the experiments is publicly available at: https://github.com/nkegke/deep-affective-bodily-expression-recognition

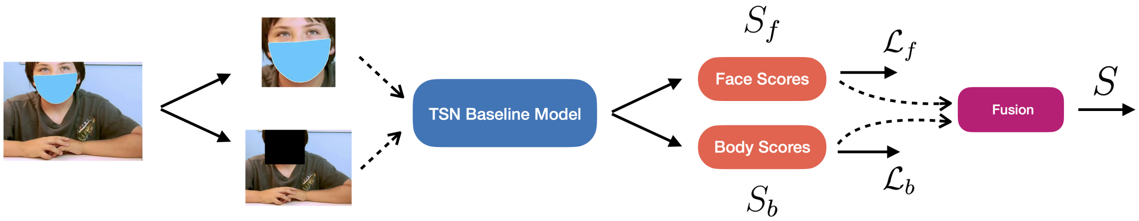


Figure 6: Modality Score Late Fusion Scheme

Table 1: TSN Training Computational Load

Segments	Time per Epoch (sec.)	
	Training	Validation
1	6	4
3	14	10
5	23	16
10	32	23

4 EXPERIMENTAL RESULTS

In this section, we present our experimental procedure and results. First, an ablation study on the number of TSN segments is performed to explore possible trade-offs. Then, we study the medical face mask effect by comparing emotion recognition results of masked input to when the faces are visible. We examine the case of when the mask is applied to the image of the full body, as well as only the face, to compare performance between input modalities. Furthermore, visual explanation techniques are utilized to display expressive features for different modalities and emotion categories. Lastly, we report results given with the enhancement techniques, both when individually utilized and when combined.

4.1 Performance vs Speed Trade-off

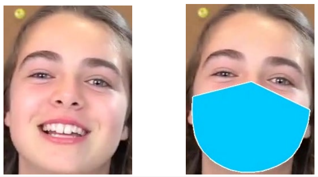
In Table 1, we perform an ablation study on the number of segments and consequently the number of snippets, which are used during the TSN training, by considering 4 different values: 1, 3, 5 and 10. By increasing the number of segments, we significantly increase computational load, and therefore inference and training time. On the other hand, when we provide the model with multiple parts of the video, it might help achieve better performance. The numbers reported stand for training with a single RTX 2080 GPU, but one could use multiple ones and increase batch size proportionally for faster training.

4.2 Mask Effect Results

We compare emotion recognition results between default and masked input, for face and full body crops. For face cropping, we extract the visual face features using OpenFace [1], an open source facial behavior analysis toolkit.

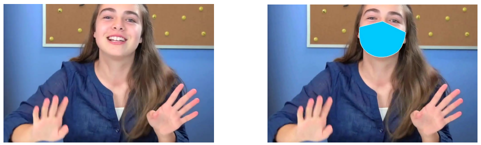
4.2.1 Mask Effect on Face Input. In Table 2, we report results on face input. At first sight, performance drops considerably (≈ 3-4%). This is a result we expected, as the mask covers the majority of the face, including one of the most expressive facial features, the mouth.

Table 2: Mask Effect Results on Face Input



Segments	ROC AUC		Performance
	Default	Mask	
1	0.755	0.728	−2.7%
3	0.769	0.733	−3.6%
5	0.767	0.732	−3.5%
10	0.770	0.741	−2.9%

Table 3: Mask Effect Results on Full Body Input



Segments	ROC AUC		Performance
	Default	Mask	
1	0.752	0.752	-
3	0.759	0.758	−0.1%
5	0.758	0.754	−0.4%
10	0.761	0.759	−0.2%

Intuitively, if one would try to predict the emotions expressed in the two images of Table 2, we sense that they would have a better chance without the presence of the mask. Regarding the number of segments used, performance peaks at 10, but increasing it above 3 does not result in significant performance difference. This means that the model does not necessarily create stronger temporal structure when provided with more than 3 parts of the video.

4.2.2 Mask Effect on Full Body Input. Looking at Table 3, which shows results on full body input, the first and most important observation we make, is that performance decrease is very little to none (0-0.4%). These results suggest that the model can exploit body information in such a way, that even with the application of a face mask, and consequently face information loss, it only suffers minimal performance drop. We also note the same pattern of performance with the masked face input results, which is better

Table 4: Masked Input Result Comparison

Segments	ROC AUC		Performance
	Masked Face	Masked Full Body	
1	0.728	0.752	+2.4%
3	0.733	**0.758**	+2.5%
5	0.732	0.754	+2.2%
10	0.741	**0.759**	+1.8%

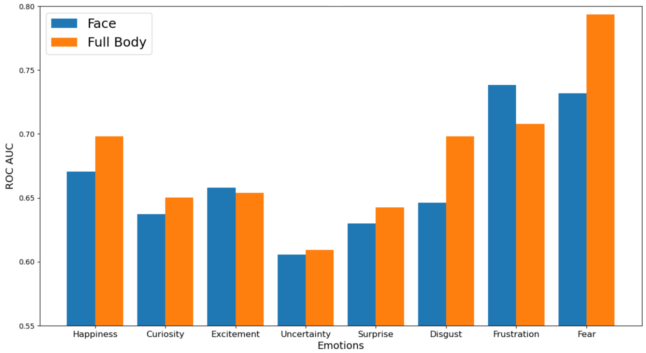


Figure 7: Masked Input Modality per Emotion Performance

performance as complexity goes up. However, performance increase from 3 to 10 segments is minimal (0.1%), which again suggests working towards the speed side of the trade-off.

4.2.3 Masked Face vs Masked Full Body Results. Lastly, in Table 4 we compare model performance with masked face versus masked full body crop, and show that incorporating the whole body in the input gives superior results over face crop. With black we highlight the best overall result, whereas with blue we highlight the result of the suggested optimal model, regarding the performance vs speed trade-off discussed earlier. The obvious conclusion is that moving towards bodily expression recognition is our best option, when the face is occluded. However, this is only a baseline result, which we could build on and pursue improvements by enhancing our model.

4.3 Per Emotion Performance

In Fig. 7, we report per emotion ROC AUC and compare masked face versus masked full body input performance. Full body outperforms face in all emotions, except for Excitement and Frustration. This could be translated as these two emotions being expressed more by facial than bodily features from the children involved and incorporating the body in this case misleads the network. For Fear, performance is a lot higher with full body compared to face, which intuitively makes sense as children tend to utilize their body more to express fear [10]. Happiness is not conventionally an emotion with intense expressions, as most people think of just a simple smile, which is obstructed by the mask, but the model manages to recognize it at a decent level. Another conclusion we could come up to is that for some emotion pairs, like Curiosity-Uncertainty or Excitement-Surprise, which intuitively are quite similar to each other, performance might be lower for each emotion individually, because it is harder for the model to distinguish one from the other.

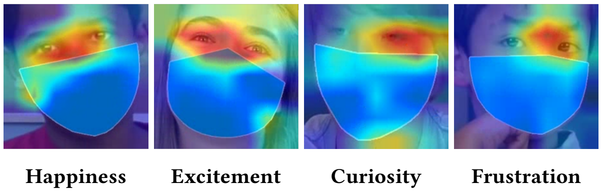


Figure 8: Face Decision Regions

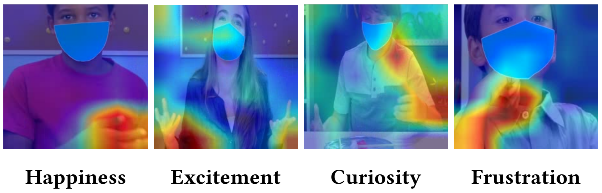


Figure 9: Body Decision Regions

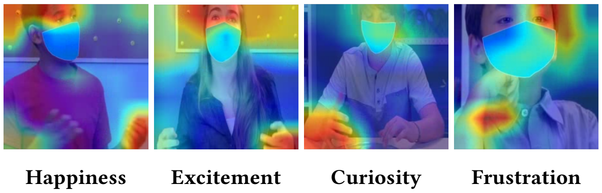


Figure 10: Mixed Decision Regions

4.4 Visual Explanation

To have a better understanding of the mask effect on performance, we utilize a technique for producing visual explanations for predictions. We wish to explore where our model focuses in the input image and how its behaviour varies for the different emotion category targets. The method we choose is Grad-CAM [33], which uses the gradients of an emotion target flowing into the final convolutional layer, to produce a coarse localization map that highlights the important regions in the image for predicting that particular emotion. We provide some example frames of the activation mapping, where model focus increases from blue to red.

Starting from the face examples (Fig. 8), we can see that the model focuses on the upper part of the face. The facial features that could be utilized are the eyes (half-closed for Curiosity, wide open for Excitement), the eyebrows (raised for Excitement), and the forehead (frowning for Frustration).

Regarding the body examples (Fig. 9), the arms become visible and provide information that is utilized by the model. The bodily features that could be utilized are the hands (calm for Happiness,

Table 5: TSN Fusion Scheme Performance Comparison

Input - 3 Segments	Aggregation	ROC AUC
Masked Face	-	0.733
Plain Body	-	0.736
Fusion	Maximum	0.724
	Average	**0.764**

Table 6: TSN vs TSM Model Performance Comparison

Model	Input - 3 Segments	Shift	ROC AUC
TSN	Masked Face	-	0.733
TSN	Masked Full Body	-	0.758
TSM	Masked Full Body	1/8	0.762
		1/4	**0.763**

Table 7: Method Combination Performance Results

Model	Input - 3 Seg.	Shift	Aggr.	ROC AUC	
				Unb.	Bal.
TSN	Masked Face	-	-	0.733	-
TSN	Masked Full Body	-	-	0.758	-
TSM	Fusion	1/8	Max.	0.729	-
			Avg.	0.767	-
		1/4	Max.	0.731	-
			Avg.	**0.768**	**0.696**
TSN	Unmasked Face	-	-	0.769	0.698
[27]		-	-	-	0.620

aroused for Excitement, investigating for Curiosity, fist for Frustration), the arms (wide open for Excitement), and the shoulders (shrugged for Curiosity).

In Fig. 10, we present several examples where the model focuses not only the body, but also on the face, fusing different modality information to make predictions. This suggests that it is able to learn both facial and bodily features in a single RGB stream.

Overall, the model has learned to ignore noisy features, like the mask and the background. It is crucial to note, that the background is considered noise in this dataset, as the videos were recorded in a directed setup and it can be the same for different reaction topics. There is also large variability introduced by differences in the children's appearance due to clothing, body shape, size and hairstyles. These examples specify that the model is able to overcome these difficulties and focus on the expressive features.

4.5 Enhancement Results

We enhance the TSN-based model with the modality fusion and temporal modeling techniques and aspire to fully overcome the consequences of the face mask, by achieving performance as high as with the unmasked input.

4.5.1 Modality Fusion. In Table 5, we report fusion results after experimenting with two different aggregation functions: maximum and average. We also present an extra row of the plain body input, the performance of which is expected to be on the same scale with the masked face. A first observation we can make is, that using maximum as the aggregation function gives poor results, as it is actually outperformed by the plain body crop method. That might happen, because we are utilizing different modality information with a single input and wrong positive predictions (false positives) from one modality are possibly canceling out correct negative predictions (true negatives) from the other. On the other hand, averaging seems a choice that blends well, as it clearly improves performance. Intuitively, it makes sense to have a balanced consensus between modalities, as emotional expression cues can vary.

4.5.2 Temporal Modeling. We experiment with the originally proposed channel shift fractions: 1/8 and 1/4. In Table 6, we observe that inserting TSM improves performance slightly. Either by shifting 1/4 or 1/8 of the channels, the difference is minimal. We come to the conclusion, that spatial feature learning plays a more important role for an emotional expression, while temporal structure is rather complementary.

4.5.3 Method Combination. In Table 7, we report results when combining the TSM and fusion techniques. It seems that when utilizing both, the same conclusions as earlier apply. That means, TSM seems to give slight temporal modeling ability to the model and the fusion method results suggest that it effectively takes advantage of the face and body information separately, and possibly avoids irrelevant information confusion. The best overall performance is 0.768 ROC AUC and is achieved by the averaging fusion method, when using TSM with 1/4 partial shift. Compared to 0.769, which is the best face result achieved with no mask applied, reported in Table 2, we almost fully overcome face information loss and achieve similar performance. For reference, the last row reports the balanced ROC AUC average result from [27], where features extracted from [1] are used with an SVM, which our TSM Fusion method clearly outperforms.

5 CONCLUSION

In this work, we studied the effect of face occlusion on a CRI visual emotion recognition problem. In the presence of a face mask, performance from just the face drops considerably and urges us to incorporate the body modality. By providing the full body image as input, the model can sustain its performance and outperform the masked face case. Spatial information can be instrumental and yield great results, while temporal structure complements fittingly, as the consensus of several video segments provides additional emotional expression information. When enhancing the baseline model with temporal modeling and more importantly modality fusion, we almost fully overcome face information loss and achieve performance similar to the unmasked input case. Our visualizations provided insights suggesting a single RGB stream can ignore noise and learn both from facial, as well as bodily expressive features. An emotion recognition system with these capabilities can effectively tackle face occlusion forced by health and safety protocols, and be a core part of various applications in crucial areas like education and health care, for both adults and children.

ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no. 101070381 (project: PILLAR-Robots).

REFERENCES

[1] Brandon Amos, Bartosz Ludwiczuk, and Mahadev Satyanarayanan. 2016. *OpenFace: A general-purpose face recognition library with mobile applications*. Technical Report. CMU-CS-16-118, CMU School of Computer Science.

[2] Hillel Aviezer, Yaacov Trope, and Alexander Todorov. 2012. Body Cues, Not Facial Expressions, Discriminate Between Intense Positive and Negative Emotions. *Science* 338, 6111 (Nov. 2012), 1225–1229. https://doi.org/10.1126/science.1224313

[3] Valentin Bazarevsky, Ivan Grishchenko, Karthik Raveendran, Tyler Zhu, Fan Zhang, and Matthias Grundmann. 2020. BlazePose: On-device Real-time Body Pose tracking. https://doi.org/10.48550/ARXIV.2006.10204

[4] Tony Belpaeme, Paul Baxter, Joachim de Greeff, James Kennedy, Robin Read, Rosemarijn Looije, Mark Neerincx, Ilaria Baroni, and Mattia Coti Zelati. 2013. Child-Robot Interaction: Perspectives and Challenges. In *Social Robotics*. Springer International Publishing, 452–459. https://doi.org/10.1007/978-3-319-02675-6_45

[5] Cynthia Breazeal. 2003. Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies* 59, 1-2 (July 2003), 119–155. https://doi.org/10.1016/s1071-5819(03)00018-1

[6] Claus-Christian Carbon. 2020. Wearing Face Masks Strongly Confuses Counterparts in Reading Emotions. *Frontiers in Psychology* 11 (Sept. 2020). https://doi.org/10.3389/fpsyg.2020.566886

[7] Claus-Christian Carbon and Martin Serrano. 2021. The Impact of Face Masks on the Emotional Reading Abilities of Children—A Lesson From a Joint School–University Project. *i-Perception* 12, 4 (July 2021), 204166952110382. https://doi.org/10.1177/20416695211038265

[8] Beatrice de Gelder. 2009. Why bodies? Twelve reasons for including bodily expressions in affective neuroscience. *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1535 (Dec. 2009), 3475–3484. https://doi.org/10.1098/rstb.2009.0190

[9] Niki Efthymiou, Panagiotis P. Filntisis, Gerasimos Potamianos, and Petros Maragos. 2021. A Robotic Edutainment Framework for Designing Child-Robot Interaction Scenarios. In *The 14th PErvasive Technologies Related to Assistive Environments Conference* (Corfu, Greece) *(PETRA 2021)*. Association for Computing Machinery, New York, NY, USA, 160–166. https://doi.org/10.1145/3453892.3458048

[10] Panagiotis P. Filntisis, Niki Efthymiou, Petros Koutras, Gerasimos Potamianos, and Petros Maragos. 2019. Fusing Body Posture With Facial Expressions for Joint Recognition of Affect in Child–Robot Interaction. *IEEE Robotics and Automation Letters* 4, 4 (Oct. 2019), 4011–4018. https://doi.org/10.1109/lra.2019.2930434

[11] Panagiotis P. Filntisis, Niki Efthymiou, Gerasimos Potamianos, and Petros Maragos. 2020. Emotion Understanding in Videos Through Body, Context, and Visual-Semantic Embedding Loss. In *Computer Vision – ECCV 2020 Workshops*, Adrien Bartoli and Andrea Fusiello (Eds.). Springer International Publishing, Cham, 747–755.

[12] Panagiotis P. Filntisis, Niki Efthymiou, Gerasimos Potamianos, and Petros Maragos. 2021. An Audiovisual Child Emotion Recognition System for Child-Robot Interaction Applications. In *2021 29th European Signal Processing Conference (EUSIPCO)*. 791–795. https://doi.org/10.23919/EUSIPCO54536.2021.9616106

[13] Goren Gordon, Samuel Spaulding, Jacqueline Kory Westlund, Jin Joo Lee, Luke Plummer, Marayna Martinez, Madhurima Das, and Cynthia Breazeal. 2016. Affective Personalization of a Social Robot Tutor for Children's Second Language Skills. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (Phoenix, Arizona) *(AAAI'16)*. AAAI Press, 3951–3957.

[14] Felix Grundmann, Kai Epstude, and Susanne Scheibe. 2021. Face masks reduce emotion-recognition accuracy and perceived closeness. *PLOS ONE* 16, 4 (April 2021), e0249792. https://doi.org/10.1371/journal.pone.0249792

[15] Hatice Gunes and Massimo Piccardi. 2007. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications* 30, 4 (Nov. 2007), 1334–1345. https://doi.org/10.1016/j.jnca.2006.09.007

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. https://doi.org/10.1109/cvpr.2016.90

[17] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A Fast Learning Algorithm for Deep Belief Nets. *Neural Comput.* 18, 7 (jul 2006), 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527

[18] Yibo Huang, Hongqian Wen, Linbo Qing, Rulong Jin, and Leiming Xiao. 2021. Emotion Recognition Based on Body and Context Fusion in the Wild. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. IEEE. https://doi.org/10.1109/iccvw54120.2021.00403

[19] Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. 2019. Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs. https://doi.org/10.48550/ARXIV.1907.06724

[20] Iolanda Leite, Ginevra Castellano, André Pereira, Carlos Martinho, and Ana Paiva. 2012. Modelling Empathic Behaviour in a Robotic Game Companion for Children: An Ethnographic Study in Real-World Settings. In *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction* (Boston, Massachusetts, USA) *(HRI '12)*. Association for Computing Machinery, New York, NY, USA, 367–374. https://doi.org/10.1145/2157689.2157811

[21] Shan Li and Weihong Deng. 2022. Deep Facial Expression Recognition: A Survey. *IEEE Transactions on Affective Computing* 13, 3 (July 2022), 1195–1215. https://doi.org/10.1109/taffc.2020.2981446

[22] Ji Lin, Chuang Gan, and Song Han. 2018. TSM: Temporal Shift Module for Efficient Video Understanding. https://doi.org/10.48550/ARXIV.1811.08383

[23] Yu Luo, Jianbo Ye, Reginald B. Adams, Jia Li, Michelle G. Newman, and James Z. Wang. 2018. ARBEE: Towards Automated Recognition of Bodily Expression of Emotion In the Wild. (2018). https://doi.org/10.48550/ARXIV.1808.09568

[24] Marco Marini, Alessandro Ansani, Fabio Paglieri, Fausto Caruana, and Marco Viola. 2021. The impact of facemasks on emotion recognition, trust attribution and re-identification. *Scientific Reports* 11, 1 (March 2021). https://doi.org/10.1038/s41598-021-84806-5

[25] Elisabeta Marinoiu, Mihai Zanfir, Vlad Olaru, and Cristian Sminchisescu. 2018. 3D Human Sensing, Action and Emotion Recognition in Robot Assisted Therapy of Children with Autism. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE. https://doi.org/10.1109/cvpr.2018.00230

[26] Dung Nguyen, Kien Nguyen, Sridha Sridharan, Afsane Ghasemi, David Dean, and Clinton Fookes. 2017. Deep Spatio-Temporal Features for Multimodal Emotion Recognition. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1215–1223. https://doi.org/10.1109/WACV.2017.140

[27] Behnaz Nojavanasghari, Tadas Baltrušaitis, Charles E. Hughes, and Louis-Philippe Morency. 2016. EmoReact: A Multimodal Approach and Dataset for Recognizing Emotional Responses in Children. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction* (Tokyo, Japan) *(ICMI '16)*. Association for Computing Machinery, New York, NY, USA, 137–144. https://doi.org/10.1145/2993148.2993168

[28] Eilidh Noyes, Josh P. Davis, Nikolay Petrov, Katie L. H. Gray, and Kay L. Ritchie. 2021. The effect of face masks and sunglasses on identity and expression recognition with super-recognizers and typical observers. *Royal Society Open Science* 8, 3 (March 2021). https://doi.org/10.1098/rsos.201169

[29] Reinhard Pekrun, Thomas Goetz, Wolfram Titz, and Raymond P. Perry. 2002. Academic Emotions in Students' Self-Regulated Learning and Achievement: A Program of Qualitative and Quantitative Research. *Educational Psychologist* 37, 2 (Jan. 2002), 91–105. https://doi.org/10.1207/s15326985ep3702_4

[30] Ioannis Pikoulis, Panagiotis P. Filntisis, and Petros Maragos. 2021. Leveraging Semantic Scene Characteristics and Multi-Stream Convolutional Architectures in a Contextual Approach for Video-Based Visual Emotion Recognition in the Wild. (2021). https://doi.org/10.48550/ARXIV.2105.07484

[31] J.C. Read and P. Markopoulos. 2013. Child-computer interaction. *International Journal of Child-Computer Interaction* 1, 1 (Jan. 2013), 2–6. https://doi.org/10.1016/j.ijcci.2012.09.001

[32] Byron Reeves and Clifford Nass. 1996. The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Pla. *Bibliovault OAI Repository, the University of Chicago Press* (01 1996).

[33] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2016. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. (2016). https://doi.org/10.48550/ARXIV.1610.02391

[34] Karen Simonyan and Andrew Zisserman. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1* (Montreal, Canada) *(NIPS'14)*. MIT Press, Cambridge, MA, USA, 568–576.

[35] Maria Tsantani, Vita Podgajecka, Katie L. H. Gray, and Richard Cook. 2022. How does the presence of a surgical face mask impair the perceived intensity of facial emotions? *PLOS ONE* 17, 1 (Jan. 2022), e0262344. https://doi.org/10.1371/journal.pone.0262344

[36] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 20–36.

[37] Sijie Yan, Yuanjun Xiong, and Dahua Lin. 2018. Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. https://doi.org/10.48550/ARXIV.1801.07455

[38] Zhengyuan Yang, Amanda Kay, Yuncheng Li, Wendi Cross, and Jiebo Luo. 2020. Pose-based Body Language Recognition for Emotion and Psychiatric Symptom Interpretation. *CoRR* abs/2011.00043 (2020). arXiv:2011.00043 https://arxiv.org/abs/2011.00043

[39] Fan Zhang, Valentin Bazarevsky, Andrey Vakunov, Andrei Tkachenka, George Sung, Chuo-Ling Chang, and Matthias Grundmann. 2020. MediaPipe Hands: On-device Real-time Hand Tracking. https://doi.org/10.48550/ARXIV.2006.10214