NITEC: Versatile Hand-Annotated Eye Contact Dataset for Ego-Vision Interaction

Thorsten Hempel, Magnus Jung, Ahmed A. Abdelrahman, and Ayoub Al-Hamadi
Neuro-Information Technology Group
Otto von Guericke University, Magdeburg, Germany
{thorsten.hempel, magnus.jung, ahmed.abdelrahman, ayoub.al-hamadi}@ovgu.de

Abstract

Eye contact is a crucial non-verbal interaction modality and plays an important role in our everyday social life. While humans are very sensitive to eye contact, the capabilities of machines to capture a person's gaze are still mediocre. We tackle this challenge and present NITEC, a hand-annotated eye contact dataset for ego-vision interaction. NITEC exceeds existing datasets for ego-vision eye contact in size and variety of demographics, social contexts, and lighting conditions, making it a valuable resource for advancing ego-vision-based eye contact research. Our extensive evaluations on NITEC demonstrate strong cross-dataset performance, emphasizing its effectiveness and adaptability in various scenarios, that allows seamless utilization to the fields of computer vision, human-computer interaction, and social robotics. We make our NITEC dataset publicly available to foster reproducibility and further exploration in the field of ego-vision interaction[1].

1. Introduction

Eye contact plays a crucial role in our everyday social interactions and is one of the most important mechanisms in non-verbal interactions [8, 18]. It serves as signal to initiative for communication [13], regulating interactions (e.g., establishing and maintaining joint attention [24, 37]) and to facilitate communication goals. The effects of eye contact among humans are also observed in human-robot interaction scenarios. In these settings, when individuals establish eye contact with a humanoid robot, it elicits similar types of automatic affective and attentional responses as they would during eye contact with another human [5, 15, 44]. Moreover, eye contact with a robot shows positive impact on its level of likability and attribution of human-likeness to a humanoid robot [19] and can even effect a humans' honesty [34].

Figure 1. Typical discussion scenario in an office, where eye contact plays a crucial role to manage the interaction. With the aid of our NITEC dataset, machines are able to reach this communication level to achieve more intuitiveness in human-machine interactions.

Humans possess a remarkable ability to perceive eye contact accurately, even in challenging conditions. However, the robust detection of eye contact in machines, particularly in the domain of human-robot interaction, has been largely unexplored, presenting a persistent and formidable challenge. We argue that one of the main reasons for this is the lack of rich and high-quality datasets to effectively train neural networks to detect eye contact robustly in unconstrained settings.

We strive to bridge this gap by introducing a new dataset, called NITEC, a dataset by the **N**euro-**I**nformation **T**echnology group for **E**ye **C**ontact detection in real life scenarios. NITEC provides hand-annotated labels for face-based eye contact detection from ego-perspective to target interaction scenarios in near- to midfield distances (*e.g.*, see Figure 1). It is based on four other datasets for different computer vision tasks, namely WIDER FACE [38], Gaze360 [14], CelebA [23], and Helen [21], that are partly re-annotated and combined to a new, well-curated dataset, that surpasses other current datasets in size, variety, and quality. In multiple experiments, we show that common CNN architectures trained on our dataset show striking

[1] https://github.com/thohemp/nitec

generalization capabilities and outstrip other models on their own datasets. Further, we analyze the qualitative performance of our baseline models (with ResNet18 and ResNet50 backbone) and study the face area and the strictness for the classification of eye contact. To summarize, our contributions are as follows:

- We introduce and publicly release NITEC, a rich and large-scale eye-contact dataset for ego-vision interaction with 36,000 hand annotated samples.
- We evaluate our dataset in numerous quantitative experiments yielding state-of-the-art results using common classification models trained on NITEC
- We conduct qualitative evaluations to gain further insights about the spatial classification behavior and its corresponding consistency, highlighting the importance of eye contact prediction models

2. Related works

There have been numerous research approaches tackling human-robot eye contact using dedicated gaze interaction systems [19, 20, 25, 29, 32, 35, 41]. Most of these systems focus on realistic robot behavior, while the human's gaze is perceived by hardware-based Eye-Trackers [33, 39] that are not applicable to real life scenarios. Other image-based approaches focus mainly on gaze vector predictions or head pose estimation to identify the current focus of attention, where eye contact can be formulated as a subtask by defining the specific gaze angle [2]. However, we will show in the following sections that gaze predictions as well as head pose estimation models are barely sufficient to fulfill this task.

Eye contact as a classification task has recently drawn significant interest in the automotive area to estimate pedestrian's attention and awareness of the traffic situation. Onkhar *et al*. [28] presented a method for deriving eye contact in traffic using a head-mounted eye-tracker for the pedestrian and an in-vehicle stereo camera. Mordan *et al*. [27] introduced an end-to-end multi-task CNN for multi-attribute pedestrian analysis, including eye-contact, based on the JAAD dataset [30]. Another dataset called "LOOK" for pedestrian eye contact detection was introduced by Belkada *et al*. [4], who proposed a body pose-based classification approach. This is caused by the fact that in the automotive application, most pedestrians are perceived from far distance, where faces alone cannot be captured with sufficient features. Smith *et al*. [36] presented one of the early works for near- to midfield ego-vision eye contact classification based on a specifically created gaze datasets. Ye *et al*. [40] presented another learning-based method, that couples a head pose-dependent appearance model with a temporal Conditional Random Field. But similar to gaze predictors, head-pose is not a reliable and precise eye contact indicator. Chong *et al*. [6] created a new image-based dataset with more than 4,000,000 samples to train neural networks that help identify the main gaze patterns for the diagnosis of Autism Spectrum Disorder. While their dataset remains non-public, Zhang *et al*. [42] and Mitsuzumi *et al*. [26] published ego-vision based eye contact annotations for existing datasets along their model proposals. However, we will show that these datasets are not sufficiently sized and qualitative enough to build robust models upon them, leaving a gap for ego-vision based eye contact detection. We strive to close this gap by introducing NITEC, a manually annotated dataset , that encompasses various scenarios, diverse environments, and different difficulty levels.

3. NITEC Dataset

In this section, we will give a detailed insight of the creation procedure and structure of our NITEC dataset. We begin with a short analysis of existing datasets, followed by details of the collection of NITEC and its annotation pipeline. Finally, we will give a short comparison of the final NITEC with other published datasets.

3.1. Existing datasets

To the best of our knowledge, there are only two publicly available datasets, that have been labeled with eye contact classes for near- and midfield detection range: DEEPEC [26] and OFDIW [42]. DEEPEC provides manually annotated images provided by the datasets LFPW [3], Helen [21], AFW [17], and IBUG [31], which sum up to 4,150 samples. The dataset is split into 53% eye contact samples and 47% samples with averted gaze. The source datasets are originally used for facial analysis and provide high-resolution, mostly non-occluded faces in unconstrained settings. OFDIW is split into an eye contact dataset for humans and eye contact dataset for animals. The human dataset consists of 16,548 samples with images collect from the LFW dataset [12], which was originally published for face recognition tasks. A third — not publicly available — dataset was introduced by Chong *et al*. [6], who conducted a study with human subjects that resulted in 4,339,879 annotated frames (281,152 with eye contact) for training and 353,924 annotated frames (25,112 with eye contact) for validation.

3.2. Data composition

Our objective was to create a comprehensive dataset for eye contact estimation in-the-wild, that provides large diversity and variability. Therefore, we selected publicly available images from four different datasets with complementary characteristics: WIDER FACE [38], Gaze360 [14], CelebA [23] and Helen [21]. WIDER FACE is a large-scale in-the-wild dataset, primarily created for face detec-

tion tasks. It contains images with varying scene context including multiple persons, where the predominantly small resolution of faces makes the classification of eye contact particularly challenging. Gaze360 is a gaze estimation dataset capturing 238 subjects in indoor and outdoor environments. In sum, it provides 172,000 samples with a wide variety of gaze directions combined with a large range of head poses. CelebA is a large-scale celebrity face dataset that focuses on face attributes. This leads to images with feature-rich faces and challenging gaze directions (e.g., slightly next to the camera). Finally, Helen is another dataset for face feature analysis without celebrities setting providing various images gathered from flicker with extraordinary range of appearance variation, including pose, lighting, expression and occlusion.

3.3. Annotation procedure

Our NITEC dataset contains in total 35,919 hand-annotated samples, where 13,829 samples are from WIDER FACE, 7,214 are from Gaze360, 12,226 are from CelebA and 2,650 are from Helen. Except for Gaze360, all samples were manually annotated using a dedicated annotation tool that provides highlighted face crops based on a RetinaFace [7] face detector. Thus, the annotators were able to incorporate the context outside the facial region into their decision-making process. The annotators were asked to subjectively decide if the selected face appears to have eye contact with the camera/annotator or not. If uncertain, faces could be skipped and excluded from the dataset. The dataset has been split into 29,003 training images and 6,916 test images, leading to a split ratio of roughly 80/20. The labeling of the training set was distributed among two annotators, while for the test set, every sample was labeled by three annotators, where the majority vote determined the final label decision. Table 1 gives an overview about the type of conflicts for each sub sets. It reports a conflict rate of roughly 15% for WIDER FACE and CelebA, and about 8% for Helen samples. Interestingly, in the latter two cases the majority vote determined in 90% an eye contact sample, while for WIDER FACE most conflicts were selected to be no-eye contact.

In contrast to the other datasets, Gaze360 provides 3D gaze vector annotations. We leverage this data, by converting the 3D gaze direction into a 2D vector, consisting of two angles *yaw* and *pitch*, and a unit vector for length. We then collected samples from the training and test set, where *yaw* and *pitch* would be between the strict thresholds of -5 and 5 degrees, indicating eye contact with the center of the camera. Likewise, we randomly sample the same number of samples with a gaze direction above the threshold for generating non-eye contact samples.

Dataset	No. of Samples	Conflicts [%]	Eye contact in conflicts [%]
NITEC-WIDER FACE	2829	15.6	24.7
NITEC-CelebA	2430	17.4	92.0
NITEC-Helen	525	8.2	88.4
NITEC	5784	15.7	59.1

Table 1. Evaluation of the mutual annotated test sets by three annotators with the number of annotated samples, the share of annotation conflicts and share of label decision based on majority vote.

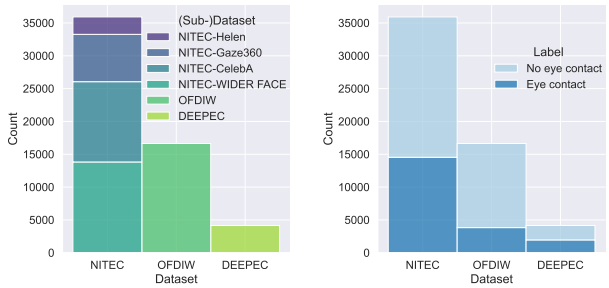

Figure 2. Comparison of our proposed dataset with two other public datasets in size and label distribution.

3.4. Dataset comparison

Our NITEC dataset will be freely available and will include the precise position of the facial region in the original image, as well as the annotations provided by the annotators. The test scripts, such as for Figure 5, will also be provided for use in research. Table 2 and Figure 2 show a comparison of our proposed dataset with the two other public datasets. With around 36,000 samples, our NITEC dataset is more than double the size of OFDIW that contains 16,648 samples. The third dataset, DEEPEC, consists only of 4,150 samples and is therefore the smallest one. However, with a label split of 47.5% eye contact samples and 52.5% it is the most balanced candidate, followed by our NITEC dataset with 40.5% eye contact. The slight overhang of non-eye contact samples is introduced by the WIDER FACE subset, where only approximately every fifth sample is labeled with eye contact, based on the subset ratio. This is caused by the nature of the WIDER FACE dataset, that contains mainly faces captured from far distances, where the person is not aware of the camera. We chose this dataset with the intention to reduce false positive in the target models for cases where target faces are feature-poor. The remaining NITEC subsets are fairly balanced around 50%. However, the OFDIW dataset has a similar label distribution compared to our WIDER FACE subset with unbalanced 23%.

Dataset	No. of Samples (Eye Contact [%])		
	Train	Test	$\sum$
OFDIW	11,511 [22.7]	4,137 [23.93]	16,648 [23.0]
DEEPEC	4,150 [47.5]		4,150 [47.5]
NITEC-WIDER FACE(Ours)	11,000 [23.2]	2,829 [20.0]	13,829 [22.6]
NITEC-CelebA (Ours)	9,829 [49.5]	2,397 [63.4]	12,226 [52.2]
NITEC-Helen (Ours)	2,125 [52.0]	525 [57.1]	2,650 [53.0]
NITEC-Gaze360 (Ours)	6,049 [50.5]	1,165 [50.6]	7,214 [50.5]
NITEC (Ours)	29,003 [39.9]	6,916 [43.0]	**35,919** [40.5]

Table 2. Comparison of our proposed dataset with other public eye contact datasets with corresponding label distribution.

4. Experiments

We conduct several experiments to analyze the performance and quality of our NITEC dataset. We begin with a quantitative analysis by comparing the performance of baseline models over multiple dataset to study the cross-dataset generalization. In a second experiment, we compare these models with other eye contact detection models and other gaze prediction and head pose based approaches. Finally, we conduct an intra-dataset experiment to evaluate the impact of each of our NITEC datasets component on the overall performance. While for our NITEC and OFDIW the train and test sets are predefined, there is no definition for DEEPEC. Therefore, we randomly split DEEPEC into 80/20 ratio for training and testing. In additional qualitative evaluations, we analyze the performance on different exemplary images and assess the prediction pattern.

4.1. Experimental setup

To train our baseline models for the binary classification between eye contact and non-eye contact, we chose the popular and simple ResNet [10] and SWIN-Transformer [22] backbone with two output neurons. The input consists of the cropped faces, and we limit the augmentation to random cropping and random horizontal flip. The model is trained for 20 epochs, using binary cross-entropy loss function with Adam optimizer [16], with a learning rate of 0.0001 (0.001 for the SWIN-Transformer) and a batch size of 80. This way, we obtain a standardized model with focus on the training data to enable optimal comparisons.

4.2. Cross-dataset evaluation

Comparing the available datasets specifically designed for eye contact detection (OFDIW, DEEPEC, and our NITEC dataset), the models were trained on the respective training sets of each dataset, and their results were compared on all test sets of the datasets and sub-datasets shown in table 3. We follow the strategy by Belkada *et al*. [4] and employ the average precision as the main metric, complemented by the F1-Score, to provide insights into the classification model's ability to accurately classify positive instances while minimizing false positives and false negatives. Examining the results of ResNet18 reveals a clear distinction between models trained on different training datasets. DEEPEC consistently performs the worst on all test datasets, showing significant differences compared to the other models (which could be attributed to the size of the training dataset). The OFDIW ResNet18 model also performs consistently worse than the NITEC ResNet18 model, even on the test data of the OFDIW dataset itself. Given the relatively poor performance of all models on the OFDIW dataset, it is likely that the OFDIW dataset is plagued by significant label noise. Analyzing the test datasets where the differences between models are most prominent, it is evident that the results on the particularly challenging WIDER FACE dataset not only perform worse compared to other test datasets but also exhibit significant variations between models, with the NITEC model consistently outperforming the others. Similarly, on the challenging CelebA dataset, which contains difficult gaze angles passing closely by the camera. As these results hold for both average precision and F1-score, the NITEC dataset enables better generalization of the relevant features for eye contact data compared to the other datasets. These findings can be extended to the more complex ResNet50 models. Here, too, the NITEC-trained model outperforms the others, except for the DEEPEC test dataset, where the model trained on DEEPEC data achieves a slightly better average precision, while the F1-score remains higher for the NITEC model. Comparing the ResNet18 models with the ResNet50 models reveals that the larger models do not yield significant improvements (except for a slight improvement in DEEPEC), but rather lose robustness. It is assumed that the ResNet50 models overfit on the datasets of this size due to faster convergence within 20 epochs.

Through a qualitative investigation, as visually depicted in Figure 3, it becomes evident that ResNet18 harnesses image information more effectively in different domains. The analysis of Figure 3 a) reveals that ResNet18 incorporates more extensive facial regions, enabling the integration of nuanced aspects such as head pose. Moreover, pertinent social cues like social smiles and diverse facial expressions find consideration within ResNet18, particularly when ocular recognition poses challenges. It is noteworthy, however, that these considerations are contingent upon the conjunction with ocular data, as evident in Figure 3 b). Conversely, ResNet50 at times exhibits an inclination to overly rely on distinct ancillary attributes, inadvertently leading to the underestimation of the significance of ocular information. The ResNet50's focus primarily centers on individual salient indicators for assessing eye contact, yet often without robust

Training Dataset	Method	Eye Contact Classification (AP) ↑ [F1-Score ↑]						
		OFDIW	DEEPEC	NITEC-WF [38]	NITEC-Gaze360 [14]	NITEC-CelebA [23]	NITEC-Helen [21]	NITEC
OFDIW [42]	ResNet18	57.4 [33.1]	70.8 [61.2]	44.3 [37.2]	76.3 [19.9]	91.6 [76.7]	92.1 [75.4]	80.4 [61.2]
DEEPEC [26]	ResNet18	31.2 [16.3]	69.6 [62.7]	27.4 [23.9]	57.8 [27.7]	80.4 [42.1]	88.6 [74.5]	62.0 [39.9]
NITEC (Ours)	ResNet18	**59.5 [55.3]**	**72.4 [73.3]**	**57.0 [59.8]**	**93.0 [86.6]**	**96.0 [90.2]**	**95.6 [89.5]**	**88.9 [83.6]**
OFDIW [42]	ResNet50	55.2 [40.0]	68.8 [59.3]	41.3 [38.7]	68.5 [19.2]	90.4 [73.7]	90.1 [72.7]	75.8 [59.0]
DEEPEC [26]	ResNet50	31.2 [10.7]	**75.7** [65.8]	26.3 [17.7]	54.3 [22.5]	83.1 [38.8]	93.0 [74.5]	63.1 [37.1]
NITEC (Ours)	ResNet50	**57.2 [53.6]**	73.8 **[71.7]**	**57.2 [57.0]**	**89.7 [84.5]**	**95.2 [90.6]**	**96.7 [90.5]**	**87.9 [83.0]**

Table 3. Comparison of different datasets using simple baseline models based on ResNet18 and ResNet50. Eye contact classification is evaluated using the average precision (AP) metric and F1-Score. Each model is trained as a simple classifier for 20 epochs on the training set of each dataset and tested on all other combinations of test sets.

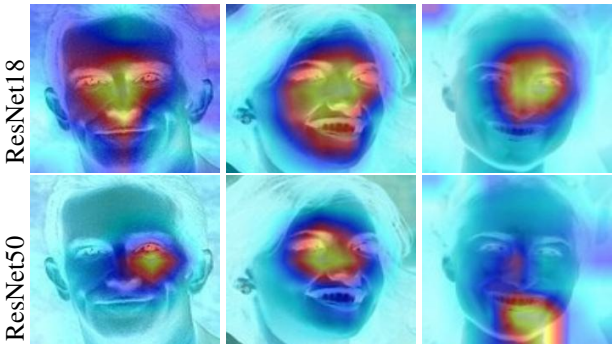

Figure 3. Visual Comparison of Gradient Class Activation Maps [9] between ResNet18 and ResNet50 on images from the CelebA dataset [23]. It is evident that ResNet18 processes more secondary information and utilizes more comprehensible image regions. (HP) refer to head pose estimation model, (G) refer to gaze estimation models.

integration with other facial features, shown in Figure 3 c). Notably, these delineated image regions pertaining to the eyes manifest a greater degree of isolation from the broader facial context, which regrettably engenders occasional fallibility in recognition, consequently resulting in the neglect of other pertinent data points.

This observation also indicates that eye contact detection tasks can be accomplished with simple architectures, highlighting the quality of the dataset, which seemingly captures relevant features for eye contact detection in various scenarios and enables efficient training.

In Table 4, we conduct another quantitative experiment. Here, instead of average precision, we measure the accuracy along with the F1 Score to include additional non-eye contact models in the comparison. For head pose-based eye contact prediction, we employ the 6DRepnet [11], a leading model for image-based head pose estimation. For the gaze direction-based approach, we utilize the L2SC-Net [1]. For further comparison, we include the Gaze360 [14] gaze direction model. Both models were trained on the Gaze360 dataset. Additionally, we include the model proposed by Chong *et al*. [6] as a comparative benchmark.

For the head pose and gaze direction models, we follow the same procedure as for the Gaze360 data labeling and define predictions as eye contact when the yaw and pitch angles are between -5 and 5 degrees. All other predictions are defined as non-eye contact. In two additional iteration, we increased this threshold to 15 and 25 degrees.

The results indicate that head pose models are not suitable for eye contact prediction. Although an accuracy of over 70 percent can be achieved on the OFDIW and WIDER FACE subsets, the accuracy should be interpreted with caution, as the label distribution for these two sets are over 70% (see also table 2). The results of the F1-score indicate that neither the recall nor the precision can achieve compatible measures. L2SC achieves similar accuracy results but reaches double-digit values in the F1-score, showing better, yet not satisfying results. The results of Gaze360 resemble those of 6DRepNet at a threshold of 5. However, when the threshold is increased to 15 or even 25 degrees, all three methods show a significant improvement for the F1-score that saturates between a threshold of 20 and 30 degrees. This can be attributed to the fact that the prediction errors of the gaze direction and head pose models are larger than the considered intervals for eye contact. This behavior is further analyzed in section 5.1.

The model by Chong *et al*. [6] was trained on the largest amount of data by far. This is evident in its consistently robust accuracy and F1-scores. While it outperforms the on OFDIW and DEEPEC trained models on the NITEC test sets, it falls behind both models on the OFDIW and DEEPEC test sets. Especially for the OFDIW datasets, the accuracy and F1-Score remains low, which supports our assumption in section 4.2 about its excessive label noise.

However, for NITEC-trained models based on ResNet and the SWIN-Transformer-tiny [22] architectures achieve

Method	Backbone	Eye Contact Classification (Accuracy) ↑ [F1-Score] ↑						
		OFDIW	DEEPEC	NITEC-WF [38]	NITEC-Gaze360 [14]	NITEC-CelebA [23]	NITEC-Helen [21]	NITEC
(HP) 6DRepNet [11]-5	ResNet50	75.4 [3.6]	54.1 [3.1]	79.6 [2.4]	50.4 [5.9]	36.4 [1.3]	43.6 [3.9]	57.0 [2.7]
(HP) 6DRepNet [11]-15	ResNet50	60.9 [28.8]	55.1 [33.5]	74.5 [22.3]	59.5 [48.2]	47.9 [42.9]	49.9 [35.7]	60.9 [39.0]
(HP) 6DRepNet [11]-25	ResNet50	50.0 [33.8]	54.7 [49.5]	70.8 [37.3]	61.0 [56.8]	61.0 [69.4]	59.0 [61.8]	64.9 [59.4]
(G) L2SC-Net [1]-5	ResNet50	72.0 [18.6]	53.4 [24.8]	79.1 [15.7]	54.5 [21.6]	46.7 [34.1]	51.0 [30.4]	61.6 [27.9]
(G) L2SC-Net [1]-15	ResNet50	59.4 [39.7]	59.4 [57.8]	74.0 [41.3]	68.6 [62.5]	66.0 [73.9]	70.3 [72.2]	70.1 [64.9]
(G) L2SC-Net [1]-25	ResNet50	46.7 [41.2]	56.7 [62.5]	66.0 [45.1]	77.3 [78.9]	69.5 [79.5]	73.9 [79.3]	69.8 [71.1]
(G) Gaze360 [14]*-5	RS18-LSTM	53.9 [11.0]	75.1 [3.7]	79.2 [4.2]	49.8 [2.3]	38.5 [9.4]	45.9 [12.3]	57.6 [7.3]
(G) Gaze360 [14]*-15	RS18-LSTM	56.2 [42.0]	68.3 [20.6]	75.3 [25.8]	57.2 [32.8]	52.9 [51.9]	57.0 [50.4]	63.1 [43.1]
(G) Gaze360 [14]*-25	RS18-LSTM	57.9 [31.3]	57.1 [57.2]	68.6 [35.9]	63.3 [54.9]	64.4 [71.8]	66.9 [71.2]	66.1 [60.7]
Chong [6]*	ResNet50	59.3 [47.8]	68.2 [45.9]	68.3 [45.9]	76.1 [73.9]	75.3 [79.9]	81.9 [83.8]	73.1 [70.2]
OFDIW [42]	ResNet18	79.3 [33.1]	67.6 [61.2]	81.7 [37.2]	54.4 [19.9]	74.8 [76.7]	76.6 [75.4]	74.3 [61.2]
OFDIW [42]	ResNet50	79.7 [40.0]	66.5 [59.3]	80.5 [38.7]	53.8 [19.2]	71.7 [73.7]	74.1 [72.7]	72.5 [59.0]
DEEPEC [26]	ResNet18	75.3 [10.7]	67.5 [62.7]	77.7 [23.9]	53.4 [27.7]	51.2 [42.1]	74.7 [74.5]	64.2 [39.9]
DEEPEC [26]	ResNet50	75.3 [10.7]	70.0 [65.8]	77.6 [17.7]	49.8 [22.5]	50.3 [38.8]	79.4 [79.5]	63.6 [37.1]
NITEC (Ours)	ResNet18	**80.6** [55.3]	**74.3 [73.3]**	**84.3** [59.8]	87.1 [86.7]	88.1 [90.3]	88.6 [89.5]	**86.4 [83.6]**
NITEC (Ours)	ResNet50	77.8 [53.6]	72.2 [71.7]	82.8 [57.0]	85.1 [84.5]	**88.3 [90.6]**	**89.3 [90.5]**	85.6 [83.0]
NITEC (Ours)	SWIN-Tiny	79.0 **[55.7]**	74.2 [71.5]	84.1 **[60.6]**	**87.8 [87.8]**	85.6 [88.1]	86.7 [87.9]	85.4 [82.6]
NITEC (Ours)	SWIN-Small	80.0 [53.9]	73.0 [70.0]	82.9 [57.4]	81.0 [78.8]	86.5 [88.7]	85.9 [87.1]	84.1 [80.4]
NITEC (Ours)	SWIN-Base	80.1 [44.1]	72.4 [67.1]	83.4 [52.8]	84.5 [83.1]	74.3 [75.3]	80.2 [79.8]	80.2 [73.0]

Table 4. Comparison of different models for eye contact classification, including head pose based and gaze based estimation methods. The used metrics are accuracy and F1-Score (in square brackets). Models with * are provided by the original authors.

the best results by a significant margin and, thus, can prevail also in this comparison as the most efficient and well-generalized model. We argue that to achieve superior results for the *small* and *base* SWIN architecture, more training data is required than the NITEC dataset currently offers.

4.3. In-dataset evaluation

Table 5 presents the results of our in-dataset evaluation of the NITEC dataset. In this evaluation, we trained the ResNet18 baseline model on the train set of each subset and tested it on all other test sets. For evaluation, we used average precision as the primary metric, supplemented with the F1-Score in parentheses. The model trained on the CelebA subset shows the strongest performance of subsets, as it outperforms not only on its own test set, but also on the WIDER FACE and Helen test set. Remarkably, the model trained on the complete dataset surpasses all other models, showcasing the synergistic effect of the composition of data from the different datasets. This highlights the complementary nature of our chosen subsets for NITEC, ultimately resulting in improved generalization performance.

	Eye Contact Classification (AP) ↑ [F1-Score] ↑				
Train/Test	WF	Gaze360	CelebA	Helen	NITEC
WF	52.8 [46.5]	75.7 [35.5]	82.1 [56.8]	87.4 [73.4]	76.6 [54.0]
Gaze360	38.8 [35.8]	87.4 [82.7]	80.5 [66.9]	84.3 [66.7]	75.5 [64.6]
CelebA	53.4 [55.0]	77.4 [52.0]	91.8 [86.0]	93.0 [86.1]	81.3 [74.1]
Helen	43.0 [39.8]	65.5 [36.7]	78.0 [61.0]	84.1 [74.9]	69.5 [54.6]
NITEC	**57.0 [59.8]**	**93.0 [86.6]**	**96.0 [90.2]**	**95.6 [89.5]**	**88.9 [83.6]**

Table 5. NITEC subset evaluation based on the ResNet18 model.

5. Qualitative analysis

For qualitative analysis, we five exemplary images and applied the ResNet-18 baseline models for NITEC, OFDIW, as well as the Chong *et al*. and the gaze-based L2SC-Net model (with a threshold of 5). The results are illustrated in Figure 4. It exemplifies that OFDIW and L2SC are incapable to detect most of the eye contact faces, while OFDIW even misclassified low-quality faces (second and third row). NITEC and Chong, however, are able to correctly determine the eye contact candidates in row two and four. Particular difference between these two models are shown by more difficult samples given in row three and five. Here, Chong predict False-Positives for heavily blurred faces in the background, while our NITEC model tends for more strict decisions. The reason for this could lie in the choice of WIDER FACE, which we selected with the intention of suppressing potential false positives in challenging images. However, in some cases this can lead to False-Negatives as shown in row five with the girl in the front. This effect is further analyzed in section 5.1.

(a) NITEC (b) Chong *et al*. [6] (c) OFDIW [42] (d) L2SC-Net [1]

Figure 4. Exemplary qualitative results of eye contact classification for NITEC, Chong, OFDIW and the gaze-based L2SC model.

5.1. Prediction distribution analysis

Figure 5 shows another qualitative comparison of eye contact/non eye contact prediction on the MPIIFaceGaze dataset [43] using baseline models by Chong *et al*. [6], Gaze360 [14], and our NITEC dataset. The MPIIFaceGaze dataset consists of 37,788 facial samples, with subjects focusing on the camera level evenly distributed within a relatively small range around the camera, excluding above the camera, resulting in a lack of information in that area. Additionally, the subjects have a similar distance from the camera. In figure 5, the predicted values are aggregated using a k-nearest-neighbors algorithm (k=100) and represented with their specific gaze target locations relative to the camera and the prediction values represented in color. The comparison includes the arithmetic mean, median, and variance. When observing the means and medians for the Chong *et al*. and our NITEC model, a downward shift of the main region classified as eye contact by the model is noticeable. However, examining the Gaze360 model reveals no such shift when considering only gaze direction. This indicates that the discrepancy lies not in the MPIIFaceGaze dataset or its evaluation but rather in the training data of the models. This can be explained by perceiving eye contact even when the whole face is observed. As the data is hand-annotated, with the eyes located in the upper third of the face, the shift occurs in the region where eye contact is detected. The graphs for the mean and median also demonstrate that the models gain more confidence as the actual focal point approaches the camera on the horizontal axis and reaches higher values when approaching just below the camera on the vertical axis. Both the Chong *et al*. model and our NITEC model exhibit a uniform decline in predicted eye contact values with increasing distance from the main eye contact region (both in mean and median). However, compared to Chong *et al*., our NITEC model is more conservative. Only 6.7% of the values predicted by the NITEC model exceed 0.75, whereas Chong *et al*.'s model has 35.7% of the values are above 0.75. On the other hand, 56.6% of the values predicted by the NITEC model are below 0.25, whereas

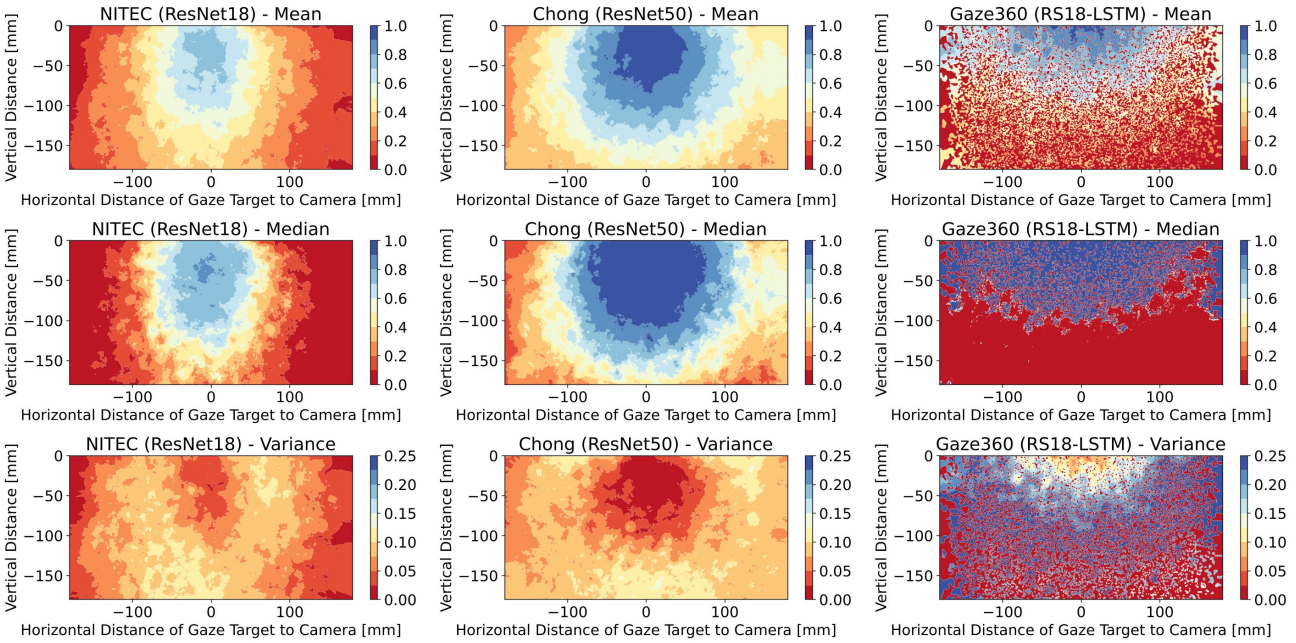

Figure 5. Qualitative comparison between different datasets using simple baseline models on the MPIIFaceGaze dataset [43]. The methods considered are Chong *et al.* [6] with a ResNet50 model, the Gaze360-LSTM trained on the Gaze360 dataset [14] with a classification threshold range of -15 to 15 degrees to qualify as eye contact, and our NITEC baseline with a ResNet18 model. For the arithmetic mean and the median, a value of 1 indicates predicted eye contact, while 0 indicates no eye contact prediction.

Chong *et al.*'s model has only 7.4% of the data points predicted below 0.25. This allows for greater adaptability of the NITEC model to practical conditions by selecting a threshold for detection. Comparing the mean and median reveals that the models decisions tend to lean towards the extremes, and the transition between eye contact and non eye contact is slower when considering average values than what the model would predict in the majority of cases, as evident in the subgraphs for the median. Another important measure for the models generalization capabilities is the scatter in predictions. Therefore, the variance within the regions derived from the k-nearest-neighbors algorithm, based on each set of 100 samples, is also shown. It can be observed that the variance is very low in the regions where the models assume eye contact (remind NITEC being more conservative than Chong *et al.*) and in the regions where no eye contact is predicted. As expected, the variance increases in the transitional areas. Overall, it becomes apparent that models trained directly on eye contact demonstrate significantly smaller variance, indicating higher robustness than the models trained solely on gaze direction. Similar to the quantitative analysis, it becomes evident that eye contact detection presents unique challenges that cannot be adequately addressed with existing gaze direction approaches. This is primarily attributed to the significant prediction error and the associated variance. This justifies the need for a dedicated dataset and specialized models for eye contact. These models offer more flexibility in adjusting the threshold to the specific application field of eye contact and provide significantly higher robustness.

6. Conclusion

In this paper, we introduced our hand-annotated NITEC dataset for image-based eye contact detection from egocentric perspective. By publicly releasing NITEC we aim to enhance research on nonverbal interaction in the field of human-machine interaction, striving to improve intuitive communication and to reduce misunderstandings. Through multiple quantitative evaluations, we have demonstrated the quality of the dataset, showcasing the exceptional generalization performance even with small baseline models. In future work, we aim to further investigate this behavior and link it with a dedicated user study to gain a better understanding of the subjective perception of eye contact.

Acknowledgments

This work is funded and supported by the Federal Ministry of Education and Research of Germany (BMBF) (AutoKoWAT-3DMAt under grant Nr. 13N16336) and German Research Foundation (DFG) under grants Al 638/13-1, Al 638/14-1 and Al 638/15-1.

References

[1] Ahmed A. Abdelrahman, Thorsten Hempel, Aly Khalifa, and Ayoub Al-Hamadi. L2cs-net: Fine-grained gaze estimation in unconstrained environments, 2022. 5, 6, 7

[2] Ahmed A. Abdelrahman, Dominykas Strazdas, Aly Khalifa, Jan Hintz, Thorsten Hempel, and Ayoub Al-Hamadi. Multimodal engagement prediction in multiperson human–robot interaction. *IEEE Access*, 10:61980–61991, 2022. 2

[3] Peter N. Belhumeur, David W. Jacobs, David J. Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, volume 35, pages 2930–2940, December 2013. 2

[4] Younes Belkada, Lorenzo Bertoni, Romain Caristan, Taylor Mordan, and Alexandre Alahi. Do pedestrians pay attention? eye contact detection in the wild, 2021. 2, 4

[5] Marwen Belkaid, Kyveli Kompatsiari, Davide De Tommaso, Ingrid Zablith, and Agnieszka Wykowska. Mutual gaze with a robot affects human neural activity and delays decision-making processes. *Science Robotics*, 6(58):eabc5044, 2021. 1

[6] Eunji Chong, Elysha Clark-Whitney, Audrey Southerland, Elizabeth Stubbs, Chanel Miller, Eliana L. Ajodan, Melanie R. Silverman, Catherine Lord, Agata Rozga, Rebecca Merrill Jones, and James M. Rehg. Detection of eye contact with deep neural networks is as accurate as human experts. *Nature Communications*, 11, 2020. 2, 5, 6, 7, 8

[7] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5203–5212, 2020. 3

[8] Teresa Farroni, Gergely Csibra, Francesca Simion, and Mark H Johnson. Eye contact detection in humans from birth. *Proceedings of the National academy of sciences*, 99(14):9602–9605, 2002. 1

[9] Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021. 5

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 4

[11] Thorsten Hempel, Ahmed A. Abdelrahman, and Ayoub Al-Hamadi. 6d rotation representation for unconstrained head pose estimation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2496–2500, 2022. 5, 6

[12] Gary B. Huang and Erik G. Learned-Miller. Labeled faces in the wild : Updates and new reporting procedures. 2014. 2

[13] Roxane J. Itier and Magali Batty. Neural bases of eye and gaze processing: The core of social cognition. *Neuroscience & Biobehavioral Reviews*, 33(6):843–863, 2009. 1

[14] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1, 2, 5, 6, 7, 8

[15] Helena Kiilavuori, Veikko Sariola, Mikko J. Peltola, and Jari K. Hietanen. Making eye contact with a robot: Psychophysiological responses to eye contact with a human and with a humanoid robot. *Biological Psychology*, 158:107989, 2021. 1

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[17] Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *IEEE International Conference on Computer Vision Workshops*, pages 2144–2151. IEEE, 2011. 2

[18] Takahiko Koike, Motofumi Sumiya, Eri Nakagawa, Shuntaro Okazaki, and Norihiro Sadato. What makes eye contact special? neural substrates of on-line mutual eye-gaze: A hyperscanning fmri study. *eneuro*, 6:ENEURO.0284–18.2019, 02 2019. 1

[19] Kyveli Kompatsiari, Francesca Ciardo, Vadim Tikhanoff, Giorgio Metta, and Agnieszka Wykowska. It's in the eyes: The engaging role of eye contact in hri. *International Journal of Social Robotics*, 13:1–11, 06 2021. 1, 2

[20] Alap Kshirsagar, Melanie Mei Hsia Lim, Shemar Christian, and Guy Hoffman. Robot gaze behaviors in human-to-robot handovers. *IEEE Robotics and Automation Letters*, 5:6552–6558, 2020. 2

[21] Vuong Le, Jonathan Brandt, Zhe L. Lin, Lubomir D. Bourdev, and Thomas S. Huang. Interactive facial feature localization. In *European Conference on Computer Vision*, 2012. 1, 2, 5, 6

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 4, 5

[23] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 1, 2, 5, 6

[24] Heidi Mauersberger, Till Kastendieck, and Ursula Hess. I looked at you, you looked at me, i smiled at you, you smiled at me—the impact of eye contact on emotional mimicry. *Frontiers in Psychology*, 13, 2022. 1

[25] Chinmaya Mishra and Gabriel Skantze. Knowing where to look: A planning-based architecture to automate the gaze behavior of social robots*. *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1201–1208, 2022. 2

[26] Yu Mitsuzumi, Atsushi Nakazawa, and Toyoaki Nishida. Deep eye contact detector: Robust eye contact bid detection using convolutional neural network. In *British Machine Vision Conference*, 2017. 2, 5, 6

[27] Taylor Mordan, Matthieu Cord, Patrick P'erez, and Alexandre Alahi. Detecting 32 pedestrian attributes for autonomous vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 23:11823–11835, 2020. 2

[28] V. Onkhar, P. Bazilinskyy, J.C.J. Stapel, D. Dodou, D. Gavrila, and J.C.F. de Winter. Towards the detection of

driver–pedestrian eye contact. *Pervasive Mob. Comput.*, 76(C), sep 2021. 2

[29] Matthew K. X. J. Pan, Sungjoon Choi, James Kennedy, Kyna McIntosh, Daniel Campos Zamora, Günter Niemeyer, Joohyung Kim, Alexis Wieland, and David L. Christensen. Realistic and interactive robot gaze. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11072–11078, 2020. 2

[30] Amir Rasouli, Iuliia Kotseruba, and John K. Tsotsos. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 206–213, 2017. 2

[31] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: database and results. *Image and Vision Computing*, 47:3–18, 2016. 300-W, the First Automatic Facial Landmark Detection in-the-Wild Challenge. 2

[32] Tetsuya Sano, Akishige Yuguchi, Gustavo Alfonso Garcia Ricardez, Jun Takamatsu, Atsushi Nakazawa, and Tsukasa Ogasawara. Evaluating imitation of human eye contact and blinking behavior using an android for human-like communication. *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1–6, 2019. 2

[33] Tetsuya Sano, Akishige Yuguchi, Gustavo Alfonso Garcia Ricardez, Jun Takamatsu, Atsushi Nakazawa, and Tsukasa Ogasawara. Evaluating imitation of human eye contact and blinking behavior using an android for human-like communication. In *2019 28th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pages 1–6, 2019. 2

[34] Elef Schellen, Francesco Bossi, and Agnieszka Wykowska. Robot gaze behavior affects honesty in human-robot interaction. *Frontiers in Artificial Intelligence*, 4, 2021. 1

[35] Shayla Sharmin, Mohammed Moshiul Hoque, S. M. Riazul Islam, Md. Fazlul Kader, and Iqbal H. Sarker. Development of duplex eye contact framework for human-robot inter communication. *IEEE Access*, 9:54435–54456, 2021. 2

[36] Brian A. Smith, Qi Yin, Steven K. Feiner, and Shree K. Nayar. Gaze locking: Passive eye contact detection for human-object interaction. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13, page 271–280, New York, NY, USA, 2013. Association for Computing Machinery. 2

[37] Sophie Wohltjen and Thalia Wheatley. Eye contact marks the rise and fall of shared attention in conversation. *Proceedings of the National Academy of Sciences*, 118(37):e2106645118, 2021. 1

[38] Shuo Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1, 2, 5, 6

[39] Zhefan Ye, Yin Li, Alireza Fathi, Yi Han, Agata Rozga, Gregory D. Abowd, and James M. Rehg. Detecting eye contact using wearable eye-tracking glasses. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, UbiComp '12, page 699–704, New York, NY, USA, 2012. Association for Computing Machinery. 2

[40] Zhefan Ye, Yin Li, Yun Liu, Chanel Bridges, Agata Rozga, and James M. Rehg. Detecting bids for eye contact using a wearable camera. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 1:1–8, 2015. 2

[41] Abolfazl Zaraki, Daniele Mazzei, Manuel Giuliani, and Danilo De Rossi. Designing and evaluating a social gaze-control system for a humanoid robot. *IEEE Transactions on Human-Machine Systems*, 44:157–168, 2014. 2

[42] Dingwen Zhang, Bo Wang, Gerong Wang, Qiang Zhang, Jiajia Zhang, Jungong Han, and Zheng You. Onfocus detection: identifying individual-camera eye contact from unconstrained images. *Science China Information Sciences*, 65, 2021. 2, 5, 6, 7

[43] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's written all over your face: Full-face appearance-based gaze estimation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*, pages 2299–2308. IEEE, 2017. 7, 8

[44] Yanxia Zhang, Jonas Beskow, and Hedvig Kjellström. Look but don't stare: Mutual gaze interaction in social robots. In *International Conference on Software Reuse*, 2017. 1