More Than Meets the Eye: Physicians' Visual Attention in the Operating Room

Sapir Gershov[1(✉)], Fadi Mahameed[1,2], Aeyal Raz[2], and Shlomi Laufer[1]

[1] Technion - Israel Institute of Technology, Technion City, 3200003 Haifa, Israel
{sapirgershov,fadi.mahamee}@campus.technion.ac.il
[2] Rambam Health Care Campus, 3109601 Haifa, Israel

Abstract. During surgery, the patient's vital signs and the field of endoscopic view are displayed on multiple screens. As a result, both surgeons' and anesthesiologists' visual attention (VA) is crucial. Moreover, the distribution of said VA and the acquisition of specific cues might directly impact patient outcomes.

Recent research utilizes portable, head-mounted eye-tracking devices to gather precise and comprehensive information. Nevertheless, these technologies are not feasible for enduring data acquisition in an operating room (OR) environment. This is particularly the case during medical emergencies.

This study presents an alternative methodology: a webcam-based gaze target prediction model. Such an approach may provide continuous visual behavioral data with minimal interference to the physicians' workflow in the OR. The proposed end-to-end framework is suitable for both standard and emergency surgeries.

In the future, such a platform may serve as a crucial component of context-aware assistive technologies in the OR.

Keywords: Eye-tracking · Surgery · Anesthesia · Visual Attention · Operation Room · Webcams · Deep Learning

1 Introduction

While under surgery, there is a multitude of clinical information about the patient that the anesthesiologist and surgeon must monitor and oversee. Because most information is presented visually, the physicians' visual attention (VA) becomes vital. Furthermore, the distribution of said VA and the acquisition of signals at specific moments during the procedure may directly impact the ability to provide better care [12,13].

Several studies investigated the preoperative team members VA [1,6,11,12,14,18] using a wearable eye-tracking device to record participants' gaze. Though these devices provide accurate data, they do not offer sustainable and ecological solutions for long-term data collection in the OR. These devices have limited

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-47076-9_2.

S. Wu et al. (Eds.): AMAI 2023, LNCS 14313, pp. 11–20, 2024.
https://doi.org/10.1007/978-3-031-47076-9_2

battery life, require a calibration stage before use, and can become inconvenient for physicians after extended usage [16]. Our study presents an alternative approach that examines the physicians' monitor observation patterns. By placing a webcam on top of the relevant monitor and continuously recording video data with minimal interference, we can recognize scenes in which the physicians' gaze is directed at the camera (i.e., direct gaze at the monitor). Such a system will facilitate the collection of vast amounts of data, enabling in-depth analysis of the medical care provider's work. Furthermore, it may serve as a crucial component of context-aware assistive technologies in the OR.

Papers in the field of eye-tracking have tackled the task of Gaze Target Prediction - detecting the attended visual target in the scene and provided new datasets, challenges, and models that produce human-like results [2,4,15,17]. These new developments have not been implemented the medical domain, though they may significantly improve medical education, training, and patient safety.

Chong et al. [2] presented a state-of-the-art architecture for detecting attended visual targets. However, this model does not distinguish between a gaze directed toward an "out-of-frame" object and a direct gaze toward the camera. This requires a modified approach that will allow the model to recognize scenes in which the physicians' gaze is directed at the camera (i.e., screen). Therefore, we chose to address this challenge using "Onfocus" detection, which identifies whether the individual's focus is on the camera [21].

Onfocus detection in unconstrained capture conditions, such as the OR, presents multiple challenges due to the complex image scenes, unavoidable occlusion, diverse face directions, constant changes in the frame focus, the number of appearing objects, and imagery factors (e.g., blur, over-exposure). Zhang et al. [21] presented a model and a dataset to evaluate onfocus detection under these challenges.

Our study presents the implementation of a webcam-based eye contact recognition model. First, we improved both Chong et al. and Zhang et al. models and provided new SOTA results. We then evaluate our model on new data - webcam videos of physicians' gaze during medical simulations (MS) and in real-life OR settings. In the future, our methodology may be employed to assess the effect of VA on patient care.

The paper's contributions are as follows: (1) an improved deep learning model for detecting the attended visual target; (2) an end-to-end eye-contact tracking framework for analyzing the distribution of VA of preoperative team members.

2 Related Work

2.1 Face Detection

YOLO ("You Only Look Once") is a popular family of real-time object detection algorithms. The original YOLO object detector was published by Redmon et al. [10]. Since then, different versions and variants of YOLO have been proposed, each providing a significant increase in performance and efficiency.

Previously, Qi et al. [8] published a modification of the YOLO architecture, YOLO5Face, which treats face detection as a general object detection task. In their work, they designed a face detector model capable of achieving state-of-the-art performance in varying image sizes by adding a five-point landmark regression head into the original architecture and using the Wing loss function [5].

2.2 Facial Landmarks

Facial landmarks (FL) detection is a computer vision task in which a model needs to predict key points representing regions or landmarks on a human's face (i.e., eyes, nose, lips, etc.).

Dlib-ml [7] is a cross-platform open-source software library with pre-trained detectors for FL. The Dlib detector estimates the location of 68 coordinates (x, y) that map the facial points on a person's face. Though newer algorithms leverage a dense "face mesh" with machine learning to infer the 3D facial surface from single camera input, these models fail to produce superior results when the acquired images have disturbances and motion [3].

2.3 Spatiotemporal Gaze Architecture

Most works that tackled the task of detecting gaze target prediction constructed 2D representations of the gaze direction, which fails to encode whether the person of interest is looking onward, backward, or sideward. Chong et al. [2] proposed to use a deep-learning network to construct a 3D gaze representation and incorporate it as an additional feature channel. The input to their network was the video frame scene, the heads positions in the frame, and the reciprocal cropped head images. However, they did not provide a face-detection model to generate this input automatically. In addition, Chong et al.'s work applied α - a learned scalar that evaluates whether the person's object of attention is inside or outside the frame, with higher values indicating in-frame attention. Yet, when the person's object of attention was outside the frame, they did not examine the cases in which the object of attention was the camera itself.

2.4 Eye-Context Interaction Inferring Network

Zhang et al. [21] provided a novel end-to-end model for onfocus detection. The model, named "Eye-Context Interaction Inferring Network" (ECIIN), is a deep learning architecture incorporating the VGG architecture for feature extraction of the eyes region and a context capsule network (CAP) [9]. Inside the ECIIN, Zhang et al. applied several network modules that implicitly explore eye-context information cues by casting the whole learning problem as an image categorization task. As it shows, Zhang et al.'s model does not take advantage of Yang et al.'s [20] publicly available face detection dataset [20], which is rich with labeled data and suitable for training a model for such a task. Furthermore, they do not

employ a well-known, state-of-the-art detection model that can be more accurate and robust using transfer learning techniques. Finally, their model was not trained for multiple object detection in the same image.

3 Materials

3.1 Benchmark Datasets

WIDER FACE Dataset [20]**.** This dataset is a subset of the publicly available WIDER dataset [19]. Currently, WIDER FACE dataset is one of the most extensive publically available datasets for face detection. It contains 32,203 images and 393,703 labels of faces with a wide range of scale, poses, and occlusion variability. Each recognizable face in the WIDER FACE dataset is labeled by bounding boxes, which must tightly contain facial landmarks (FL) (e.g., forehead, chin, and cheek). In the case of occlusion, the face is labeled with an estimated bounding box.

VideoAttentionTarget Dataset [2]**.** This dataset was created specifically for video gaze target modeling and accommodated 1,331 annotated videos of people's dynamic gaze behavior in diverse situations. The videos, which were gathered from YouTube, are of various domains, and they were trimmed to contain dynamic gaze behavior in which a person of interest can be continuously observed. A trimmed video duration ranges from 1–80 s.

OFDIW Dataset [21]**.** The dataset has several unique characteristics: (1) the dataset contains videos of individuals during face-to-camera communication; (2) the data is collected from a single camera point-of-view, and most of the recorded faces are completely visible; (3) the camera is focused on the presented individual where there is very little change between frames. The dataset comprises 20,623 unconstrained images with good age diversity, facial characteristics, and rich interactions with surrounding objects and background scenes. Therefore, while the OFDIW dataset provides a great starting point, it lacks a few components crucial for onfcous detection in the OR settings. For that reason, we created our unique datasets.

3.2 Our Datasets

Our datasets consist of webcam video recordings from three different setups - one from medical simulations (MS) and two from real-life OR settings (See Fig. 1). The first OR dataset focuses on anesthesiologists' work (See Fig. 1-B), and the second on the surgeon's work during minimally invasive thoracic surgery (See Fig. 1-C).

Our simulations combine two main components: a high-fidelity manikin that mimics the human body and its physiological responses and a patient monitor that presents the patient's vital signs. Data includes 31 simulations with 33 residents, and the setup was located inside the hospital's post-anesthesia care

Fig. 1. Illustration of the images in our dataset. (A) Simulated patient monitor point-of-view. The participants' faces are manually blurred; (B) Real patient monitor point-of-view; (C) Thoracic surgery setup. All shown subjects have given their consent to have their pictures featured.

unit (PACU), using real-life medical equipment and utilities. To document the resident's visual patterns during MS, we collected data using a single webcam above the patient monitor (See Fig. 1-A). The simulations dataset contains 31 videos, each approximately 20 min long.

Over 1200 frames of MS have been extracted from the videos. The appearing faces were manually labeled with bounding boxes suitable for the YOLO network, and an independent human observer marked the events of direct eye contact with the monitor. These events are classified as "Onfocus" or "Out of focus" while maintaining a class-balanced and diverse dataset (see Fig. 2).

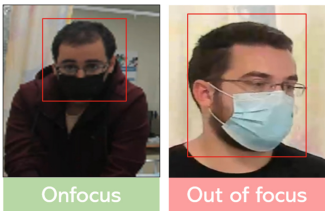

Fig. 2. The onfocus detection task labels. All shown subjects have given their consent to have their pictures featured.

To capture the anesthesiologist VA, we placed the webcams above two OR monitors - a patient monitor and a ventilator monitor. Ten different anesthesiologists (6 male, 4 female) were recorded during 11 surgeries.

Four thoracic surgeons (all male) were recorded by placing the webcams on top of a screen tower (See Fig. 1 - C). A board-certified thoracic surgeon and a resident executed each surgery. The OR dataset contains 11 videos, each approximately 2 h long.

The OR datasets comprise 1873 frames (1473 frames of anesthesiologists and 400 frames of surgeons) from the available OR videos. These frames were manually labeled as described for the MS dataset (see Fig. 2). Once again, we made sure that the dataset was balanced.

The Institutional Review Board of Rambam Health Care Campus approved the study.

4 Methods

4.1 Pipeline Construction

For face detection, we trained YOLOv7 on the WIDER FACE dataset [20]. For each detected bounding box prediction, an FL algorithm was applied. We used only the coordinates visible in the collected data - eyes, nose, and mouth.

The coupling of YOLOv7 trained for face detection with a FL detector is a suitable replacement for Zhang et al. [21] ECIIN-designed network modules. Our modifications harness the benefits of well-trained object detectors and large datasets and thus produce superior results. Then, once the region of interest is located, we apply the process described in Zhang et al. work to generate the onfocus detection.

Lastly, we modified Chong et al. work by adding the ECIIN. This modification has improved Chong et al. model performance in cases where the object of attention is "out-of-frame".

The complete end-to-end pipeline is depicted in Fig. 3.

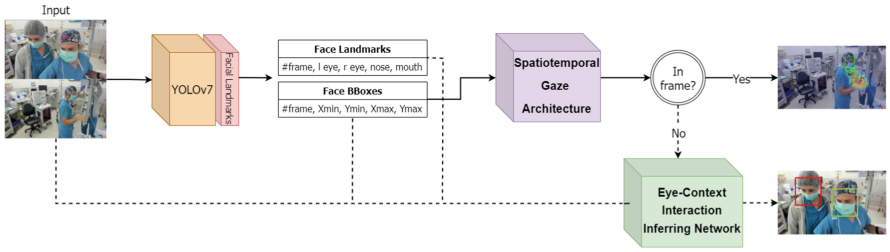

Fig. 3. End-to-end framework pipeline. The Spatiotemporal prediction is indicated by a bounding box over the allocated head and a heatmap over the object of attention. The ECIIN network classification confidence is indicated by color, where green is for high confidence and red is for low confidence. The score next to the bounding box is the prediction probability. (Color figure online)

4.2 Implementation

The training of the YOLOv7 for face detection was executed using AdamW optimizer with an initial learning rate of 1E-3 and weight decay of 5E-3. The training procedure ran for 250 epochs with a batch size of 64, and the loss was calculated using the YOLO loss function. To avoid overfitting, we stopped training when the validation loss increased.

The ECIIN and the Spatiotemporal model have been fine-tuned based on their provided source code. The trained YOLOv7 was fine-tuned by training it

for another 60 epochs. Again we stopped the training process when the generalization error increased.

The models are implemented in PyTorch 1.7.1 and trained using 2 NVIDIA RTX A6000 GPUs.

5 Results

5.1 Ablation Study of the End-to-End Pipeline

The following results are generated by utilizing the published models' weights and without fine-tuning the hyper-parameters:

Table 1. Results of the proposed modifications on different benchmarks and comparison to other architectures

Model	Task	Dataset	Accuracy	F1-Score
YOLO5Face [8]	Face Detection	WIDER Face	86.55%	-
YOLOv7 & FL			**87.03%**	0.85
ECIIN [21]	Onfocus Detection	OFDIW	84.71%	0.90
ECIIN [21] with YOLOv7 & FL			**84.97%**	0.90
Spatiotemporal [2]	Gaze Target Prediction	VideoAttentionTarget	86.12%	0.85
Complete pipeline			**87.2%**	**0.86**

5.2 Evaluation of the Proposed Framework on Our Datasets

We fine-tuned the different models using the MS dataset and evaluated their performance on our labeled OR frames. We applied 5-fold cross-validation, dividing the dataset into 70% train, 10% validation, and 20% test. Throughout this procedure, we avoided using the same videos for both the training and testing of the model.

To assess the Spatiotemporal model [2] onfocus detection performance, we used our YOLOv7 model for face detection and addressed the Spatiotemporal model predictions as binary classification (i.e., the object of attention is inside the frame or outside) (Table 2).

Table 2. Models onfocus detection results on MS frames.

Model	Train Dataset	Test Dataset	Accuracy	F1-Score
ECIIN	MS	MS	63.98% ±2.53%	0.64 ±0.03
Spatiotemporal			71.03% ±1.87%	0.72 ±0.11
Complete pipeline			**89.22% ± 1.26%**	**0.87 ± 0.02**

Table 3 results are generated by testing the fine-tuned models on the OR datasets.

Table 3. Models onfocus detection results on real OR frames.

Model	Train Dataset	Test Dataset	Accuracy	F1-Score
ECIIN	MS	OR-Anesthesiologists	52.44%	0.55
Spatiotemporal			76.44%	0.77
Complete pipeline			**86.43%**	**0.87**
ECIIN	MS	OR-Surgeons	61.19%	0.59
Spatiotemporal			75.88%	0.73
Complete pipeline			**90.01%**	**0.90**

6 Discussion

In the field of surgery and anesthesiology, eye tracking became a popular methodology for investigating the visual behavior of physicians in their natural environment. However, most eye-tracking technology is intrusive, interferes with the participants' natural workflow, and is unsuitable for prolonged data collecting.

Therefore, we employ webcams, which are considered non-intrusive, and provide continuous visual behavior data in real-time without interfering with the medical personnel workflow. This work presents an end-to-end pipeline for processing and analyzing raw video recordings of preoperative team members' workflow in real-life OR settings. The data was collected via two webcams inside the OR and later processed with a deep-learning model for gaze target prediction. The first step in the pipeline is to locate the masked faces in a frame and, for each detected face, to extract the eyes region. To do so, we harnessed the potential of YOLOv7, a state-of-the-art object detection model, and a well-trained model for FL detection. After the faces and eye regions are allocated, Chong et al. [2] model deduces if the object of attention is located in the frame based on the α scalar value. In the last stage of the pipeline, we applied Zhang et al.'s [21] model for onfocus detection. These modifications have proven fruitful in making our model more robust and accurate than the original models (See Table 1) and suitable for real-life OR settings (See Table 3). In addition, our approach is not limited by the number of participants or their distance from the cameras.

It is important to note that although the complete pipeline was not fine-tuned using the public datasets it was tested on, it achieved better results. Furthermore, the complete pipeline has a significantly faster average inference speed, at 23 FPS, compared to other models. However, more thorough research is required to quantify the influence of each component on the framework performance.

We recognize that our small-scaled datasets are a limitation of this study. Indeed, further work is required to fully explore the applications of gaze target prediction in the OR settings. Another significant limitation is the camera field

of view. There have been occasions in which the camera has not captured the participants while they had a clear view of the screen. This limitation can be overcome by adding more cameras to watch over different areas.

Prospect of Application: This unique approach for eye-tracking in a real-life OR setting may provide fresh and essential insight into surgery, anesthesia, and other fields. In the future, we trust that our non-intrusive framework could lay the groundwork for using gaze patterns, specifically onfocus detection, as an early alarm system to reduce clinical errors and as a metric to assess VA. In addition, gaze target prediction may facilitate developing an empiric metric for investigating medical personnel VA and evaluating its impact on patient outcomes.

References

1. Chetwood, A.S.A., et al.: Collaborative eye tracking: a potential training tool in laparoscopic surgery. Surgical Endoscopy **26**(7), 2003–9 (2012). https://doi.org/10.1007/s00464-011-2143-x. http://www.ncbi.nlm.nih.gov/pubmed/22258302
2. Chong, E., Wang, Y., Ruiz, N., Rehg, J.M.: Detecting attended visual targets in video. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 5395–5405 (2020). https://doi.org/10.1109/CVPR42600.2020.00544. https://github.com/ejcgt/attention-target-detection
3. Deng, J., Guo, J., Ververas, E., Kotsia, I., Zafeiriou, S.: Retinaface: single-shot multi-level face localisation in the wild. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 5202–5211 (2020). https://doi.org/10.1109/CVPR42600.2020.00525
4. Fang, Y., et al.: Dual attention guided gaze target detection in the wild. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 11385–11394 (2021). https://doi.org/10.1109/CVPR46437.2021.01123
5. Feng, Z.H., Kittler, J., Awais, M., Huber, P., Wu, X.J.: Wing loss for robust facial landmark localisation with convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2235–2245 (2018)
6. Gil, A.M., Birdi, S., Kishibe, T., Grantcharov, T.P.: Eye tracking use in surgical research: a systematic review. J. Surg. Res. **279**, 774–787 (2022). https://doi.org/10.1016/j.jss.2022.05.024. https://linkinghub.elsevier.com/retrieve/pii/S0022480422003419
7. King, D.E.: Dlib-ml: a machine learning toolkit. J. Mach. Learn. Res. **10**, 1755–1758 (2009)
8. Qi, D., Tan, W., Yao, Q., Liu, J.: YOLO5Face: Why Reinventing a Face Detector (2021). https://www.github.com/deepcam-cn/yolov5-face. http://arxiv.org/abs/2105.12931
9. Ramasinghe, S., Athuraliya, C.D., Khan, S.H.: A context-aware capsule network for multi-label classification. In: Leal-Taixé, L., Roth, S. (eds.) ECCV 2018. LNCS, vol. 11131, pp. 546–554. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11015-4_40

10. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016-Decem, pp. 779–788 (2016). https://doi.org/10.1109/CVPR.2016.91. http://pjreddie.com/yolo/
11. Roche, T.R., et al.: Anesthesia personnel's visual attention regarding patient monitoring in simulated non-critical and critical situations, an eye-tracking study. BMC Anesthesiology **22**(1) (2022). https://doi.org/10.1186/s12871-022-01705-6. https://doi.org/10.1186/s12871-022-01705-6
12. Schulz, C.M., et al.: Visual attention of anaesthetists during simulated critical incidents. British J. Anaesthesia **106**(6), 807–813 (2011). https://doi.org/10.1093/bja/aer087. www.anvil-software.de
13. Szulewski, A., Egan, R., Gegenfurtner, A., Howes, D., Dashi, G., McGraw, N.C., Hall, A.K., Dagnone, D., Van Merrienboer, J.J.: A new way to look at simulation-based assessment: the relationship between gaze-tracking and exam performance. Canadian J. Emergency Med. **21**(1), 129–137 (2019). https://doi.org/10.1017/cem.2018.391
14. Tien, T., Pucher, P.H., Sodergren, M.H., Sriskandarajah, K., Yang, G.Z., Darzi, A.: Eye tracking for skills assessment and training: a systematic review. J. Surgical Res. **191**(1), 169–178 (2014). https://doi.org/10.1016/j.jss.2014.04.032. https://linkinghub.elsevier.com/retrieve/pii/S0022480414004326
15. Tomas, H., et al.: GOO: a dataset for gaze object prediction in retail environments. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 3119–3127 (2021). https://doi.org/10.1109/CVPRW53098.2021.00349. https://github.com/upeee/GOO-GAZE2021
16. Wagner, M., et al.: Video-based reflection on neonatal interventions during COVID-19 using eye-tracking glasses: an observational study. Archives of disease in childhood. Fetal Neonatal Edition **107**(2), 156–160 (2022). https://doi.org/10.1136/archdischild-2021-321806. https://fn.bmj.com/content/107/2/156 https://fn.bmj.com/content/107/2/156.abstract
17. Wang, B., Hu, T., Li, B., Chen, X., Zhang, Z.: GaTector: a unified framework for gaze object prediction. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 19566–19575. IEEE, June 2022. https://doi.org/10.1109/CVPR52688.2022.01898. https://ieeexplore.ieee.org/document/9879784/
18. White, M.R., et al.: Getting inside the expert's head: an analysis of physician cognitive processes during trauma resuscitations. Ann. Emerg. Med. **72**(3), 289–298 (2018). https://doi.org/10.1016/j.annemergmed.2018.03.005
19. Xiong, Y., Zhu, K., Lin, D., Tang, X.: Recognize complex events from static images by fusing deep channels. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 07–12-June, pp. 1600–1609 (2015). https://doi.org/10.1109/CVPR.2015.7298768
20. Yang, S., Luo, P., Loy, C.C., Tang, X.: WIDER FACE: A face detection benchmark. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 2016-Decem, pp. 5525–5533 (2016). https://doi.org/10.1109/CVPR.2016.596, http://mmlab.ie.cuhk.edu.hk/projects/
21. Zhang, D., Wang, B., Wang, G., Zhang, Q., Zhang, J., Han, J., You, Z.: Onfocus detection: identifying individual-camera eye contact from unconstrained images. Science China Inf. Sci. **65**(6), 1–12 (2022). https://doi.org/10.1007/s11432-020-3181-9