

Video object detection for privacy-preserving patient monitoring in intensive care

Raphael Emberger*, Jens Michael Boss†, Daniel Baumann†, Marko Seric†, Shufan Huo†‡, Lukas Tuggener*, Emanuela Keller†, Thilo Stadelmann*§

{embe, tugg, stdm}@zhaw.ch, {firstnames.surname}@usz.ch

*Centre for Artificial Intelligence, ZHAW School of Engineering, Winterthur, Switzerland

†Neurocritical Care Unit, Department of Neurosurgery and Institute of Intensive Care Medicine, Clinical Neuroscience Center, University Hospital Zurich and University of Zurich, Switzerland

‡Neurology, Charité - University Medicine Berlin, Berlin, Germany

§European Centre for Living Technology (ECLT), Ca' Bottacin, Venice, Italy

Abstract—Patient monitoring in intensive care units, although assisted by biosensors, needs continuous supervision of staff. To reduce the burden on staff members, IT infrastructures are built to record monitoring data and develop clinical decision support systems. These systems, however, are vulnerable to artifacts (e.g. muscle movement due to ongoing treatment), which are often indistinguishable from real and potentially dangerous signals. Video recordings could facilitate the reliable classification of biosignals using object detection (OD) methods to find sources of unwanted artifacts. Due to privacy restrictions, only blurred videos can be stored, which severely impairs the possibility to detect clinically relevant events such as interventions or changes in patient status with standard OD methods. Hence, new kinds of approaches are necessary that exploit every kind of available information due to the reduced information content of blurred footage and that are at the same time easily implementable within the IT infrastructure of a normal hospital. In this paper, we propose a new method for exploiting information in the temporal succession of video frames. To be efficiently implementable using off-the-shelf object detectors that comply with given hardware constraints, we repurpose the image color channels to account for temporal consistency, leading to an improved detection rate of the object classes. Our method outperforms a standard YOLOv5 baseline model by +1.7% mAP@.5 while also training over ten times faster on our proprietary dataset. We conclude that this approach has shown effectiveness in the preliminary experiments and holds potential for more general video OD in the future.

Index Terms—object recognition, medical informatics, DCAI

I. INTRODUCTION

The intensive care unit (ICU) is a challenging work environment, which demands high staffing and constant alertness toward emergencies. Numbers, curves, and alarms from multiple medical devices, although well intended, often cause additional stress. As a consequence, severe burnout syndrome is present in about 50% of critical care physicians and one-third of critical care nurses [1], which in turn has been shown to correlate strongly with intent to seek other career opportunities. This is particularly problematic as the healthcare labor shortage has been exacerbated since the onset of the Covid-19 pandemic in many European countries.

To reduce the burden on healthcare professionals and physicians, clinical decision support systems, and early warning systems promise to assist healthcare professionals in decision-

making and outcome prediction. Thus avoiding cognitive overload and consequent treatment errors. These systems take advantage of the vast number of biosignals generated at high resolution by patient monitors and other medical devices. In a neurocritical care setting, these biosignals include for example arterial pressure, intracranial pressure, blood, and brain tissue oxygenation, electrocardiography, and electroencephalography recordings. Despite the crucial role these biosignals play in clinical patient assessment, the clinical implementation of machine learning solutions taking full advantage of them is hindered by artifacts in the signals as well as a lack of context in which the signals were acquired [2]. Artifacts can for example be caused by patient motion or staff interventions. However, without appropriate contextual knowledge, it is not possible to correctly interpret biosignals and distinguish physiological features from artifacts, even though many domain-specific signal processing techniques have been developed [3].

To address this challenge and to gain access to contextual information, we have implemented a camera monitoring system to detect the presence of patients and staff members, thus laying the foundation for more accurate artifact removal approaches. Ultimately resulting in better decision support systems and thus better patient outcomes, while decreasing the burden on clinical staff by false alarms. However, the system is subject to the following constraints: (a) To respect the privacy of patients, staff, and visitors, video footage can only be stored severely blurred (see Figure 1), removing most of the visual cues to detect relevant objects. (b) Also for privacy reasons, the system has to run on-site on hospital hardware resulting in narrow computational constraints.

In this paper, we propose a video OD method to address the preceding challenges, enabling privacy-preserving patient monitoring in clinical practice. Specifically, our contribution is the extension of a lightweight off-the-shelf still image OD method (that can efficiently run on standard hardware) to learn from the temporal succession of video frames without architectural changes (such that implementation and integration can be performed efficiently). Experimental evaluation shows +1.7% improvement in mean average precision with 0.5 IoU overlap while training over ten times faster than the baseline.

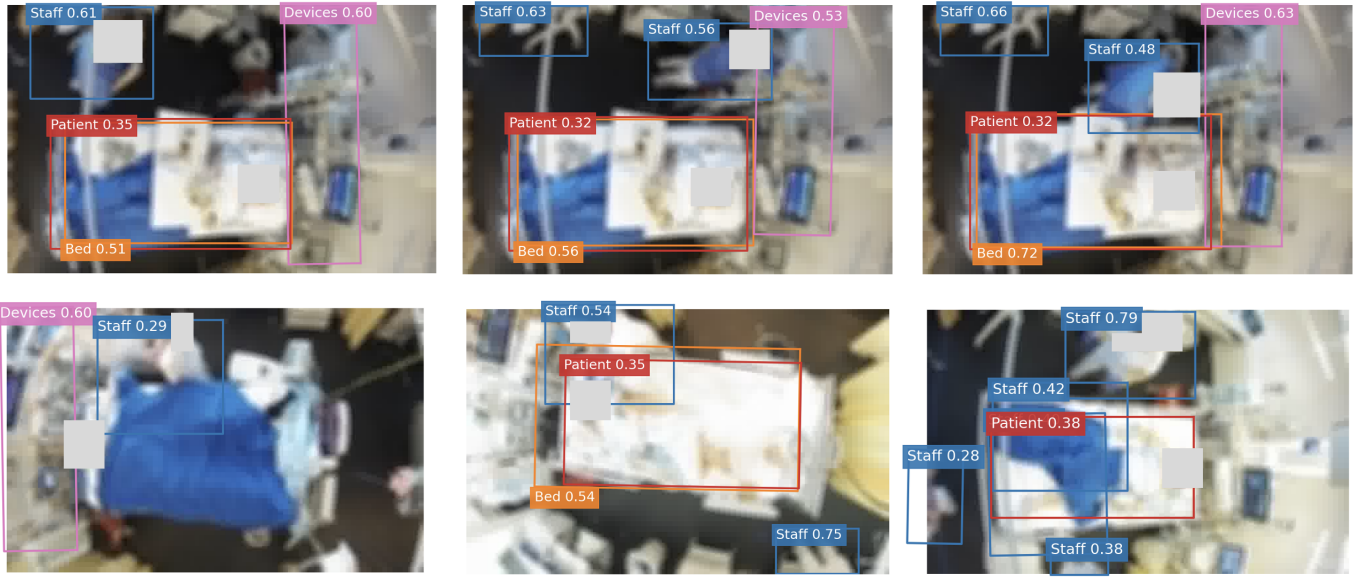


Fig. 1. Examples from our dataset with overlaid detections of the proposed model. *Top row*: Successive frames a typical scenario. In the first frame (*left*), all objects are correctly detected. *Middle* and *right* frames show how the healthcare staff is tracked moving along the bed to a medical device and then to the patient’s bedside. A white stand at the top left is wrongly identified as staff. *Second row*: Out-of-context examples where even human observers have difficulty recognizing objects. *Left*, the patient is completely covered by a blue blanket, occluding the patient and bed. *Middle*, the lighting is comparatively bright, making it difficult for the model to detect the medical devices. *Right*, the blue blanket was confused with a staff member’s blue jacket.

II. RELATED WORK

Camera-based patient motion detection for measuring vital signs and false alarm reduction has been studied in many contexts employing optical flow and artificial neural networks, as well as using 3D cameras [4]–[6]. However, even though simple motion quantification approaches showed promising results, they do not provide the same amount of context information as OD methods [7], which yield time-resolved position and class of objects visible in each frame. OD for patient monitoring has not been studied widely in the literature. Existing studies have incompatible prerequisites like high-resolution video input [8], or employ off-the-shelf OD methods [9] on unblurred data. The quasi-industry standard for state-of-the-art OD method (also used in [9]) is the YOLO family of models [10], [11]. Of these, the YOLOv5 variant [12] is closest to the application context of this work, as it allows for oriented bounding boxes [13] and is considered one of the best performing, easy to use and lightweight models.

From a still image of the clips in our dataset, it is hard even for humans to pick out where members of staff are in a frame, due to the high blur and lack of context (cp. Figure 1). However, once the same frame is seen in the context of a video, the motions are made visible to viewers and it is easier to identify members of staff. YOLOV [14] can leverage chronological video frames stacked together as one sample, exploiting motion information in our dataset. However, preliminary experiments have shown that this model is not suitable for this application. Also, domain adaptation techniques to leverage larger pretrained models [15] have been trialed, but rejected after preliminary experiments.

III. A METHOD FOR PRIVACY-PRESERVING VIDEO OD

For the baseline detector, YOLOv5 with oriented bounding boxes [13] is chosen for its simplicity and wide usage in practical settings. To enable this lightweight model to do what humans do—exploit the temporal consistency of video frames and the information induced by motion—we add information on the last frame into the current one. We encode this additional information in the existing RGB channels:

The *red channel* is replaced by a grayscale representation of the original image. Even though the gray-scaled image is harder to interpret even for humans, the general shape of the objects in the picture is still intact. Therefore, the filters of the convolutional layers that are applied to this channel would still be able to detect shapes. This decision was influenced by the lack of information in color concerning the object classes.

The *green channel* is repurposed to represent large pixel changes in comparison to the previous frame. Hence, the channel indicates movement to the model if detected to encourage learning to distinguish object classes that move more frequently (e.g. staff members) from those that do not (like the typically stationary patients). This simplistic pixel change indication is only applicable because the video is observing a relatively still environment, and the camera does not change its angle, position, or perspective.

The *blue channel* is replaced by a bitmap that contains the area of the previous frame’s bounding boxes (either ground truth during training or predictions during inference), marked by an arbitrary value (32), and otherwise 0. This encourages the model to consider the bounding boxes from the predictions or the ground truth of the last frame for the current frame.

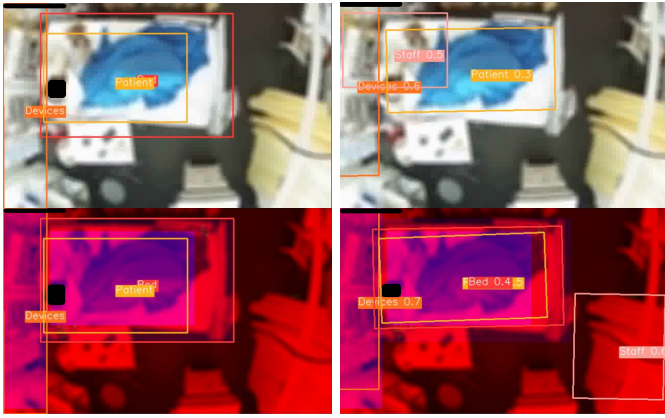


Fig. 2. OD ground truth (left) and the predictions (right) on a sample image. The top row shows the baseline model in- and output, and the bottom row the proposed model (grayscale image in red channel).

To prevent an overreliance on this channel, only a randomly chosen half of the samples have non-zero bitmaps in the third channel. Of the half that has bounding box information from the previous frame, 20% were randomly chosen to be discarded completely to account for new objects appearing in the image, as well as for missing detections from previous frames. Furthermore, 60% of the bounding boxes’ areas were randomly moved around up to 10 pixels according to a uniform distribution to account for minor local variations in the earlier frame’s predictions, just as object movement.

With this repurposing of the RGB channels, we provide the model with useful information about the temporal consistency of the frame succession (and thus object movement) without having to adapt YOLOv5’s efficiently executable and well-proven architecture, thus making development convenient. We hypothesize that this gain in temporal context information more than compensates for the loss of color information through the reduction of the current frame to grayscale, leading to higher OD rates in the experiments.

IV. EXPERIMENTAL SETUP AND RESULTS

Data Collection For the development of the OD models, we prospectively collected blurred anonymized video data from cameras (AXIS M1065-L) directed onto the bedsides of a 12-bed neurocritical care unit at the University Hospital Zurich. The blurred video streams have a resolution of 640×400 pixels at 25 frames per second, collected by a dedicated research IT infrastructure [16]. The video data streams are blurred using a software solution for video stream conversion (FFmpeg, <https://ffmpeg.org/>) with a box blur filter (`boxblur=6:1`). The blurring is required to ensure the privacy of clinical and hospital staff members as well as visitors of patients. Written informed consent was received by all patients or by their legal representatives. The study (part of the project “ICU Cockpit”) was approved by the ethics committee of Kanton Zurich (Basec no. 2021-01089), Switzerland, and was conducted following the ethical standards of the 2013 declaration of Helsinki for research involving human subjects.

Predicted (baseline)	Ground Truth				
	Bed	Staff	Devices	Patient	BG
Bed	0.31	0	0	0.02	0.05
Staff	0	0.71	0	0	0.83
Devices	0	0	0.93	0	0.01
Patient	0.04	0	0	0.69	0.11
BG	0.65	0.29	0.07	0.3	0
Predicted (Ours)	Ground Truth				
	Bed	Staff	Devices	Patient	BG
Bed	0.95	0	0	0.08	0.01
Staff	0	0.75	0	0	0.89
Devices	0	0	1	0	0.01
Patient	0.02	0	0	0.82	0.09
BG	0.04	0.25	0	0.1	0

Fig. 3. Confusion matrices for the baseline (top) and proposed model (bottom). Note that “BG” means “Background”.

Model	Epochs	mAP@.5(%)				
		Bed	Staff	Devices	Patient	All
YOLOv5	119	98.0	58.1	97.6	95.3	87.2
Proposed	10	99.5	58.1	98.4	99.4	88.9

TABLE I
AMOUNT OF RECEIVED TRAINING (EPOCHS) AND MAP@.5 FOR THE BASELINE AND PROPOSED MODEL BY CLASS AND OVERALL.

Data Preprocessing and Labeling Videos are collected as 24-hour recordings ensuring the capturing of different lighting conditions as well as situations from night and day shifts. The 24-hour recordings are cut such that individual videos can be attributed to individual patients; only videos from patients who have given informed consent are kept. To extract time periods that show movement automatically, pixel-wise differences between frames are calculated and used as a metric for general motion. By choosing an appropriate threshold, we can identify video snippets that included medical personnel with a high probability. An additional 10 s of video data is added to either side of the identified video snippets, resulting in videos that typically range from one to several minutes in length. Through this process, in total 30.748 clips are accumulated. From these clips, 196 clips are selected to be hand-labeled, with a balanced representation of the different beds and scenarios. The chosen labels are “patient”, “bed”, “staff”, and the location of the (medical) “devices”—as the corresponding objects are the ones that play the most crucial role in establishing context for the situation at a patient’s bedside. The bounding boxes of the labels can be rotated as well; e.g. the bed frame and the patient laying within, or staff walking around, turning, or leaning over the bed and patient.

Clinical Compute Infrastructure Due to the handling of patient-related data, hardware options for model training are severely restricted. The only compliant option is a virtual machine with 8 CPUs (2 GHz), 32 GB RAM, 500 GB SSD storage, and two NVIDIA TITAN V GPUs. For deployment, the final OD model would run on a server without GPUs.

Baseline and Training Details The YOLOv5 “s”-version is chosen for both the baseline and proposed method as it is the second smallest one and therefore very attractive to be deployed into the production environment without GPUs. To train the models, the dataset is split on a frame-by-frame basis into 70% training data, 15% validation data, and 15% test data. Both models were trained until they did not show any improvement for 100 consecutive epochs up to maximum 300 epochs. They are trained using the default hyperparameters provided by YOLOv5 (0.01 learning rate and 0.937 momentum after a warm-up period of three epochs; before that, 0.1 learning rate and 0.8 momentum). The model used an L2 regularization (weight decay) of 0.0005.

Results Models are evaluated using the mean average precision (mAP) metric at 0.5 IoU overlap as well as the amount of training measured in epochs. As shown in Table I, the training of the baseline model lasted for 119 epochs (after which it showed no further improvement), while the proposed method required only a fraction (8.4%) of this training time. As seen in both the confusion matrix and the example outputs in Figures 2 and 3 respectively, the baseline model has trouble recognizing members of staff and the bed frame. While our proposed model also struggles with the staff category, it does significantly better on bed frames and patients. When comparing the mean average precision at 0.5 IoU overlap, the proposed method outperforms the baseline by 1.7% (refer to Table I).

V. DISCUSSION AND CONCLUSIONS

We demonstrated that our proposed method outperforms the baseline model using only a fraction of the training time. We attribute this improvement to the new data format in which information is presented to the model in the repurposed image channels, enabling it to learn from temporal correlations. While it loses the color information, the additional information about pixel changes and bounding boxes from the earlier frame more than compensates for this loss, evidenced by the performance increase and shortened training time. What seems like a hack is typical for deep learning in practice: In absence of large training sets and conditions as found in public benchmarks [17], the available information has to be exploited optimally while considering computational boundary conditions.

We identify great potential in exploring the proposed method further, including not replacing the RGB channels, but instead expanding them with additional channels. There is also the question of the efficacy of this method when not using stationary videos, but using changes in perspective, viewing angle, etc. However, we hypothesize that it will increase the efficiency of training of any given model, and leave respective

experiments to future work, together with necessary ablation studies w.r.t. hyperparameters that did not fit the scope of this short communication. Note that the proposed method is model-independent and applicable to architectures beyond YOLOv5.

From a medical AI perspective, we anticipate that even the limited contextual information extracted by the proposed method can contribute significantly to improved artifact detection and handling. For example, staff presence can now be used during a preprocessing step or directly as input to other machine-learning models to reduce false alarms due to wrong measurements. A different application would be to determine the level of care received by individual patients to optimize the assignment of nursing staff and anticipate possible situations of nursing overload.

Acknowledgement This work has been supported by DIZH grant “AUTODIDACT”, the SNSF, and the CSB Berlin.

REFERENCES

- [1] N. Embriaco et al., “High level of burnout in intensivists: Prevalence and associated factors,” *Am. J. Respir.*, vol. 175, no. 7, pp. 686–692, 2007.
- [2] I. Elezi, *Exploiting contextual information with deep neural networks*. PhD thesis, Ca’Foscari University of Venice, 2020.
- [3] M. K. Islam, A. Rastegarnia, and S. Sanei, *Signal Artifacts and Techniques for Artifacts and Noise Removal*, pp. 23–79. Cham: Springer International Publishing, 2021.
- [4] C. Muroi, S. Meier, V. De Luca, D. J. Mack, C. Strässle, P. Schwab, W. Karlen, and E. Keller, “Automated False Alarm Reduction in a Real-Life Intensive Care Setting Using Motion Detection,” *Neurocrit Care*, vol. 32, no. 2, pp. 419–426, 2020.
- [5] Y. S. Dosso, S. Aziz, S. Nizami, K. Greenwood, J. Harrold, and J. R. Green, “Video-Based Neonatal Motion Detection,” *Proc. EMBS*, vol. 2020–July, pp. 6135–6138, 2020.
- [6] C. Coronel et al., “3D Camera and Pulse Oximeter for Respiratory Events Detection,” *IEEE J Biomed Health Inform.*, vol. 25, no. 1, pp. 181–188, 2021.
- [7] L. Jiao, R. Zhang, F. Liu, S. Yang, B. Hou, L. Li, and X. Tang, “New generation deep learning for video object detection: A survey,” *IEEE Trans Neural Netw Learn Syst.*, 2021.
- [8] S. S. Abdul Rajjak and A. K. Kureshi, “Recent advances in object detection and tracking for high resolution video: Overview and state-of-the-art,” in *Proc. ICCUBEA*, pp. 1–9, 2019.
- [9] M. A. Gul, M. H. Yousaf, S. Nawaz, Z. Ur Rehman, and H. Kim, “Patient monitoring by abnormal human activity recognition based on cnn architecture,” *Electronics*, vol. 9, no. 12, 2020.
- [10] J. Redmon et al., “You only look once: Unified, real-time object detection,” in *Proc. CVPR*, pp. 779–788, 2016.
- [11] T. Diwan, G. Anirudh, and J. V. Tembhurne, “Object detection using YOLO: Challenges, architectural successors, datasets and applications,” *Multimed Tools Appl.*, vol. 82, pp. 9243–9275, 2022.
- [12] G. Jocher et al., “Ultralytics/yolov5: v7.0 - YOLOv5 SotA realtime instance segmentation.” DOI 10.5281/ZENODO.3908559, 2022.
- [13] Hukaixuan19970627, sunbuhui, and Ethan-niu, “YOLOv5 for oriented object detection.” Available online: https://github.com/hukaixuan19970627/yolov5_obb, 2021. Last accessed: Feb 27, 2023.
- [14] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, “Yolox: Exceeding yolo series in 2021,” 2021.
- [15] P. Sager, S. Salzmann, F. Burn, and T. Stadelmann, “Unsupervised domain adaptation for vertebrae detection and identification in 3D CT volumes using a domain sanity loss,” *J Imaging*, vol. 8, no. 8, 2022.
- [16] J. M. Boss et al., “ICU Cockpit: a platform for collecting multimodal waveform data, AI-based computational disease modeling and real-time decision support in the intensive care unit,” *JAMIA*, vol. 29, no. 7, pp. 1286–1291, 2022.
- [17] T. Stadelmann, V. Tolkachev, B. Sick, J. Stampfli, and O. Dürr, “Beyond ImageNet: deep learning in industrial practice,” in *Applied data science*, pp. 205–232, Springer, 2019.