IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 January 2023, accepted 27 January 2023, date of publication 1 February 2023, date of current version 8 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3241835

RESEARCH ARTICLE

Contactless Drink Intake Monitoring Using Depth Data

RACHEL COHEN[1,2], GEOFF FERNIE[1,2,3], AND ATENA ROSHAN FEKR[1,2]

[1]KITE—Toronto Rehabilitation Institute, UHN, Toronto, ON M5G 2A2, Canada
[2]Institute of Biomedical Engineering, University of Toronto, Toronto, ON M5S 3G9, Canada
[3]Department of Surgery, University of Toronto, Toronto, ON M5T 1P5, Canada

Corresponding author: Rachel Cohen (rachelj.cohen@mail.utoronto.ca)

This work was supported by the Canadian Institutes of Health Research (CIHR) Foundation under Grant FDN-148450.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the KITE-TRI-UHN Research Ethics Board under Application No. 21-5132 on August 21, 2021.

ABSTRACT It is important for humans to remain hydrated, particularly for older adults who are at a greater risk of dehydration and may forget to drink. Monitoring liquid intake and getting reminders to drink throughout the day is a useful solution to increase hydration levels. The objective of this paper is to automatically detect drink events from multiple containers in a simulated home environment using a vision-based approach. The proposed work compares the use of depth and RGB (red, green, blue) cameras for this task. In this paper, we compared 2D and 3D Convolutional Neural Networks (CNN) using RGB and depth cameras. We collected data from nine participants performing drinking, eating and other Activities of Daily Living (ADL) in a simulated home environment. We found that for the 3D models, the RGB and depth camera inputs provided very similar F1-scores for both 10-Fold (94.3% vs 93.9%, respectively) and Leave-One-Subject-Out (LOSO) cross validation (84.2% vs 86.2%, respectively). This is a promising result as depth cameras also mitigate the challenges to privacy of RGB-based models. The 3D CNN models outperformed the 2D models, thereby creating a more robust system. Depth cameras are a useful alternative to RGB cameras with equal performance in identifying drinking events.

INDEX TERMS Artificial neural networks, computer vision, depth cameras, fluid intake monitoring, image recognition, intake gesture detection, video signal processing.

I. INTRODUCTION

Remaining hydrated is an important factor of being healthy, especially as we age. Unfortunately, many older adults do not consume enough liquid to stay hydrated, leading to adverse consequences such as hospitalization or even death. Several factors increase the changes of dehydration as we age. Our sense of thirst diminishes, and we often forget to drink enough [1], [2]. Other bodily changes such as decreased water content and reduced kidney function also increase our likelihood and severity of dehydration as we age [1]. As our population continues to age, more older adults will wish to age in their homes independently. Creating tools to monitor and prevent dehydration in a home environment will become increasingly useful to prevent hospitalizations.

The associate editor coordinating the review of this manuscript and approving it for publication was Wenbing Zhao.

Getting reminders to drink throughout the day can be beneficial to increase overall hydration. Therefore, there is a need for a system that tracks the amount of liquid consumed and sends reminders to drink only when needed. The first step of this process is to detect when a drink occurs, which is the main objective of this paper.

Other systems have attempted to track hydration using wearables or sensors in the bottle, however each of these have their own pitfalls. Wearables must be recharged regularly and one must remember to put them on. Sensors in containers also need recharging, and limit which container you can drink out of. Having an ambient system that requires no input from the user is an attractive alternative.

In our previous work, we analyzed if RGB camera data (i.e. normal cameras) could classify drinking events in a home environment [3]. This manuscript expands on this by comparing the use of depth camera data for this task. The

motivation of this manuscript is to determine if using depth cameras is a viable option to detect drink intake events.

Depth data are an attractive option compared to normal cameras, as they have better potential to preserve privacy, a key factor when deploying cameras in home settings. They are also influenced less by different lighting conditions. Instead of capturing an image of the environment, a depth camera captures the distance of each pixel to the camera and creates a depth map. Though more privacy preserving, depth data provide less detailed information and might have noise in the signals. This paper analyses if using data from depth cameras and deep learning can accurately detect drinking events in a simulated home environment, comparing it to models trained on RGB data.

In this paper, we also compare the use of 2D CNN models to 3D CNN models. 2D models take in a single image as the input and attempt to classify each individual frame, whereas 3D models receive multiple frames as input (i.e small videos) and classify the entire group of frames at once. We hypothesized that using 3D inputs with depth images would have a superior performance, as it is classifying motion of the drink which is easier to see in depth images compared to classifying drinking in single frames. To the best of our knowledge, no other papers have used depth videos to detect drink intake.

II. PREVIOUS WORK

Previous work has investigated multiple ways to detect a drink event, mainly including wearables, containers or vision-based approaches [4]. This section will focus on outlining the methods that have previously used depth data to classify drink events, which have all used the Microsoft Kinect to capture RGB and depth frames. Chua et al. used only the depth images of a Microsoft Kinect to extract hand grasping postures using Haar-like features in the frame [5]. This could detect whether a hand was grasping a cup with an 88% true positive rate (F1-score of 84.3%) [5]. Tham et al. was one of the few to exclusively use depth images from a Kinect with dynamic time warping to classify drinking among other activities with an F1-score of 93% [6]. Chang et al. showed that these two could be combined to first detect the hand holding the cup, then detect the activity such as drinking or spoon holding [7]. Due to the common problem of occlusion with vision based approaches, Cippitell et al. used a top down orientation with the Kinect placed on the ceiling to capture a meal intake [8], [9], [10]. Since the Microsoft Kinect does not function properly with a top-down orientation, they developed algorithms using a Self-Organized Map algorithm to track the person's movements. They achieved 98.3% accuracy to detect drinking [8]. This was tested on 35 participants during a meal. However, the dataset was limited, as it only included images of mealtime gestures. This work was expanded in [9] to improve the real-time capabilities which created a more automatic identification of the first and last frame of a food intake event. Cunha et al. placed a Kinect in front of 3 elderly users consuming a meal. Based on the joint coordinates, mealtime intake events were detected with an average success rate of 89% [11]. Costa et al. extended this study by also adding Hidden Markov Methods to classify the events [12]. They found that different methods were better at classifying liquid than food intake, and left-handed liquid intake proved to be a challenge [12]. Kassim et al. used a Kinect in front of the person to determined intake events during a meal to predict the calories consumed. They achieved an overall accuracy of 96%, however it was only tested on one subject [13]. Hondori et al. fused a Kinect with a wrist inertial sensor to detect eating and drinking, however this was only a pilot study with 1 participant [14].

FIGURE 1. HomeLab layout at KITE Research Institute. The yellow box represents the areas in the field of view of the cameras, the red dots represent the locations of the RGB cameras and the blue dot represents the depth camera.

Of the papers these papers listed, all used classification of static frames (2D) to detect drinking. Rouast et al. as well as our previous work using RGB signals shows that using multiple frames for drink recognition can yield better results compared to individual frames [3], [15]. Rouast et al. previously showed this with meal-time events, comparing multiple 3D deep learning architectures to 2D deep learning architectures, achieving an accuracy of. This paper builds upon these by comparing the use of a depth camera to an RGB camera. To the best of our knowledge, no study has attempted to detect fluid intake events using deep learning with multi-frame inputs of depth frames. Therefore, the main contribution of this manuscript is to show if 3D CNN can better detect drink intake events with depth cameras than traditional models used in previous works. This will be potentially used as a privacy preserving, accurate solution for ambient drink intake monitoring.

III. METHODS

A. DATA COLLECTION

The experiments were performed in HomeLab, a simulated home laboratory, at the KITE Research Institute, TRI-UHN (Figure 1). Ethical approval for the study was obtained from the KITE-TRI-UHN Research Ethics Board (21-5132) on August 21, 2021, and participants gave written informed consent prior to study participation.

Nine subjects, 5 male and 4 female with an average age of 24±3 (Mean±Standard deviation), performed all

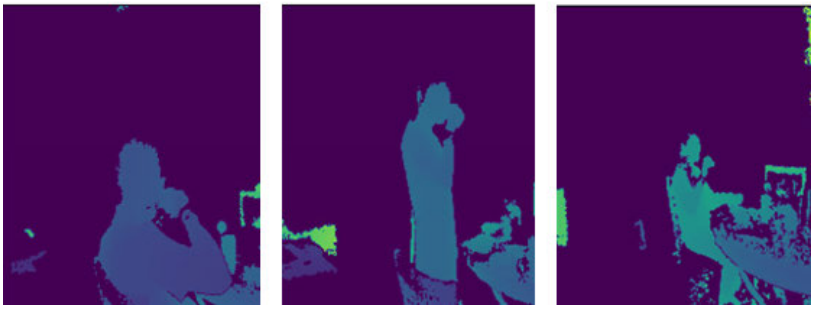

FIGURE 2. Examples of the drinking events using the depth camera.

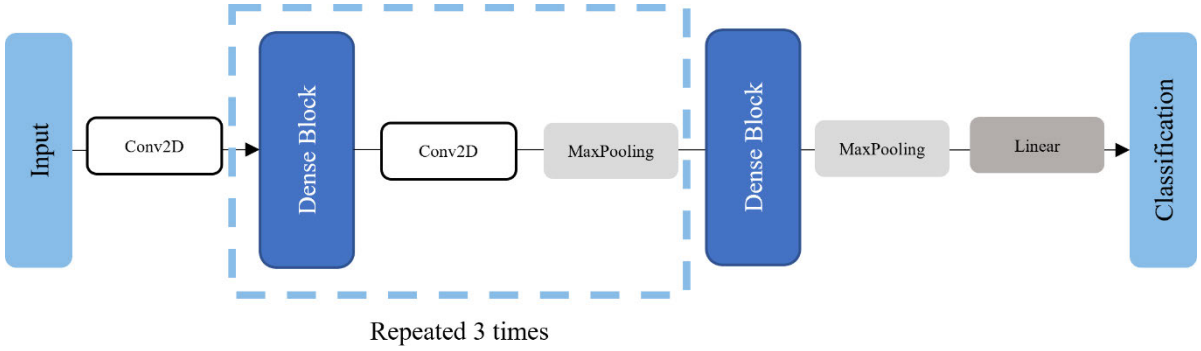

FIGURE 3. Architecture of a 2D CNN model, the DenseNet121 which performed well in most hyperparameter combinations.

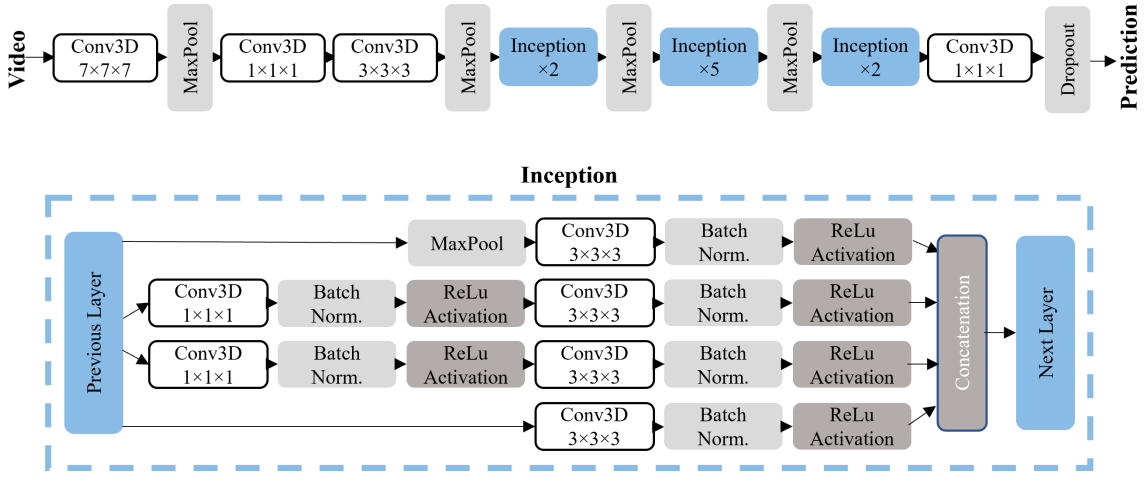

FIGURE 4. Architecture of the 3D CNN model, proposed by Carreira et al. [16].

experiments. The experiments included a controlled drinking scenario and other activities. The controlled drinking scenario had two parts: (1) where the subject drank a small, medium, or large sip based on their perception and comfort, and (2) where the subject drank a single or double sip. A double sip consists of the subject swallowing twice without placing the container to the ground. In the activity scenario, the subjects drank as much as they wanted, ate with a spoon, fork and their hands and performed other activities of daily living around the space (Table 1). The experiments were performed in a random order and the order of the containers, the location of the subject, and the sip size were also all randomized. In each experiment, the subject drank a total of 12 sips using their dominant hand from the 12 containers (Table 1). The eating events were repeated three times each for each subject.

B. DATA EXTRACTION AND PRE-PROCESSING

Videos were collected from an RGB and a depth camera, simultaneously. The events were labeled by both the subject and the researcher pressing a button when the vessel touched the lips of the subject to collect the ground truth. In this paper, an Intel RealSense L515 Lidar camera was used which has an embedded RGB camera. This camera has a field of view of $70° \times 55°$ ($\pm 3°$) with frequency of 30 fps. The maximum resolution is 1024×768.

TABLE 1. List of all activities performed in our study.

Drinking Containers	Activities of Daily Living
• A teacup with hot liquid • A coffee mug with hot liquid • Two metal commercial water bottles • A plastic, disposable water bottle • A can • A glass tumbler with ice water • A glass tumbler with colored liquid (pop or juice) • A plastic, short colored cup with water • A wine glass with colored liquid (pop or juice) • Two glass tumblers with a straw	• Scratching their head and face • Pointing the TV remote and watching TV • Doing their hair/touching their head • Using a laptop • Using a smartphone (calling and texting) • Pouring water from a kettle • Stretching • Washing the counters • Putting on and taking off a jacket • Walking around the apartment • Talking to the researcher • Writing • Folding laundry • Eating with fork, spoon, and hands 3x each

For the multi-frame input models, hereafter referred to as the 3D model, the videos were downsampled to 6fps and 3fps. Videos of 3sec and 10sec were extracted to train the model. If the timestamp indicated by the ground truth was within the window of the frames the entire video data was labeled as a drink event. The two frame rates and window sizes were evaluated to determine if they had an impact on the performance of the model. For the single-frame input, hereafter referred to as the 2D model, the exact frames in the 10sec videos inputs were used. Only the frames where the vessel was directly touching the lips of the subject were labeled as drinks. This allowed us to fairly compare the 2D and 3D models, as they were analyzing the same frames.

As the participants were in a simulated home environment, the background was complex and we had previously found that sometimes this background confused the model [3]. To mitigate this, for the depth data it was possible to remove the foreground and background based on the depth values. Then, the noise including small contours were also removed based on the size of the connected component after erosion (Figure 2).

C. NEURAL NETWORK CONFIGURATIONS

In this paper, different CNN models were trained to perform a binary classification identifying drinking and non-drinking events. Using either RGB or depth data, a 2D CNN was used to classify individual frames and 3D CNN was used to classify multi-frame inputs. We used Transfer Learning to build both 2D and 3D models. Transfer learning is a technique in which we use previous architecture already trained on datasets containing large amounts of data in a similar application and then transfer this knowledge to the model with our own dataset. This technique reduces the amount of data required and, in many cases, increases the accuracy as compared to models built from scratch.

Multiple parameters and hyperparameters were tested to find the best combination. This included

- Frame rate: 3 and 6 fps were tested for the 3D models as the former requires less computation to train and test, while the later provides more information per input.
- Window size: 3 and 10 second window inputs were tested for the 3D models. Ten seconds was chosen as the entire drink fits in one input, while 3 seconds contains partial drinks.
- Hyperparameters: were adjusted such as batch size and learning rate
- Number of layers trained: Either all of the layers or only the top layer (known as feature extraction) were trained
- Sampling method: To overcome the class imbalance, various methods such as class weights, undersampling and oversampling were compared
- Pre-trained models: For the 2D models, 8 state-of-the-art pre-trained models were tested. This includes DenseNet169, DenseNet121, InceptionResNetV2, InceptionV3, Xception, MobileNetV2, NasNetLarge, ResNet. These are all commonly used, state-of-the-art models. The model that most commonly performed the best amongst the different hyperparameter combinations is shown in Figure 3. For the 3D data, only one pretrained architecture was chosen called Inflated 3D Conv Nets (I3D) proposed by Carreira et al., which is based on the Inception V1 architecture but with inflated layers to allow 3D (video) inputs [16]. It achieved a 71.1% accuracy on the Kinetics dataset which includes 400 human activities obtained from YouTube videos. It outperformed other 3D deep learning models on benchmark datasets such as ImageNet. We also added a dropout layer to reduce overfitting. This model's architecture can be found in Figure 4.

Additionally, we applied Fine Tuning to the models where the last 2/3rds of the layers were retrained and refined. We applied the early stopping mechanism to reduce overfitting.

Out of all the combinations listed above, the model that yielded the highest F1-score was chosen and is presented in this paper. F1-score was chosen as the key metric since the two classes were imbalanced, so it properly displays the balance between precision and recall.

IV. RESULTS AND DISCUSSION

The best models for each category were the model that yielded the highest F1-score. We tested both 10-Fold and Leave-One-Subject-Out (LOSO) cross validation. The 3D models outperform the 2D models, particularly for LOSO for both depth and RGB data (Table 2).

LOSO validation is a more conservative approach as it has less biases and is a more realistic approach to what would happen in a real-world deployment. The 2D inputs contain the exact frames as the 3D models in order to properly compare them. This led to some images that were similar to each other, as they are close together in time and the participant may have had little movement between frames. This can introduce bias into the test data particularly in 10-fold validation; therefore

TABLE 2. Performance metrics for the top models. the values in the brackets represent the standard deviations for each.

Model	Validation	Sampling	Desc. Hyperparam.	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
3D	10 Fold	RGB	3 s, 3fps, class weights, 32 batch, feature extraction	97.1 (±0.6)	98.3 (±1.2)	90.7 (±2.3)	94.3 (±1.2)
		DEPTH	3 s, 3 fps, class weights, 16 batch	96.2 (±4.4)	96.9 (±5.8)	92.5 (±12.6)	93.9 (±8.1)
	LOSO	RGB	10 s, 3 fps, class weights, 16 batch	88.6 (±9.4)	88.6 (±12.5)	85.8 (±20.7)	84.2 (±14.8)
		DEPTH	3 s, 3 fps, class weights, 16 batch, feature extraction	94.8 (±7.3)	97.1 (±5.2)	84.0 (±30.0)	86.2 (±23.2)
2D	10 Fold	RGB	DenseNet121, class weights, 32 batch	98.7 (±0.6)	91.4 (±5.4)	96.2 (±3.5)	93.7 (±3.3)
		DEPTH	DenseNet121, undersample, 32 batch	88.0 (±22.4)	82.3 (±35.0)	82.5 (±34.9)	78.2 (±34.4)
	LOSO	RGB	Xception, undersample, 16 batch, feature extraction	95.7 (±1.0)	83.2 (±6.4)	70.0 (±20.7)	73.6 (±14.3)
		DEPTH	DenseNet121, undersample, 32 batch, feature extraction	80.6 (±10.7)	60.8 (±19.0)	68.2 (±21.0)	62.3 (±15.5)

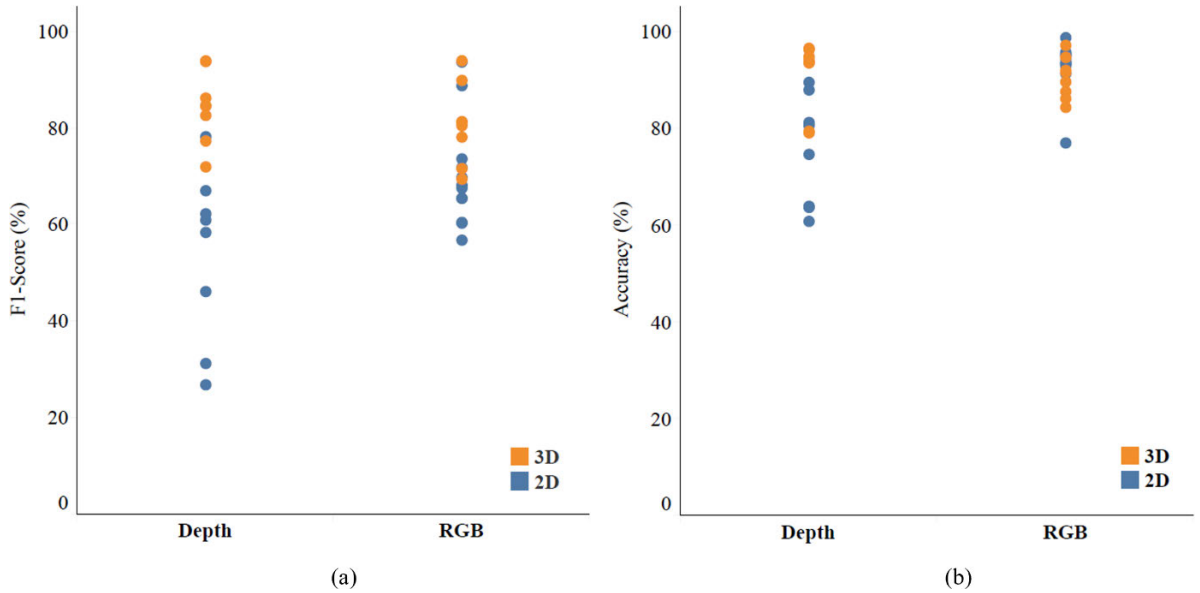

FIGURE 5. Comparison of (a) F1-Score and (b) Accuracy for trained models of different hyperparameter combinations. The 3D data is shown in orange whereas the 2D is in blue.

it is more appropriate to look at the LOSO validation for the 2D models.

In the 3D models, the depth and RGB inputs had almost the same results. The depth LOSO results achieved a slightly higher F1-score (86.2%) compared to RGB (84.2%), but with a larger standard deviation. This shows that using depth inputs may be an equally viable solution for this task with multi-frame inputs. Depth cameras are privacy preserving, as the individual is not identifiable, making them advantageous.

For both depth and RGB inputs, the 2D models were inferior to the 3D models. Figure 5 shows the F1-Scores and Accuracy values of all tested models. This confirms that 3D models are more robust at identifying these activities for both RGB and depth inputs. Interestingly, when analyzing the 2D models, the depth inputs underperformed compared to the RGB for both 10-Fold and LOSO cross validation (62.3% vs 73.6% F1-score). This is in-line with the previous literature, which has shown that depth cameras generally performed worse than RGB cameras. However, our paper shows that although this is true for the traditional 2D models, using 3D models trained on depth data creates a model with equivalent performance to the models trained by RGB frames.

Confusion matrices (Figure 6) indicate that, the 3D model trained on depth inputs had more false negatives than false positives for the LOSO validation whereas the 3D model trained on RGB data provided larger false positive rates. Please note that, the reason the number of samples varies across the confusion matrices is that we selected the best models for each scenario (ex. Number of frames per sample, type of sampling etc.). In our case, false positives are more critical as they are predicting that the person drank when they did not, which could overestimate the amount consumed and not prompt the user, leading to a dehydration status. Therefore, considering this criterion, the 3D model trained on depth is outperforming the model trained on RGB data. However, in our application, equal amounts of false negatives and false positives are preferable since throughout multiple sips, the overall error could balance out. The confusion

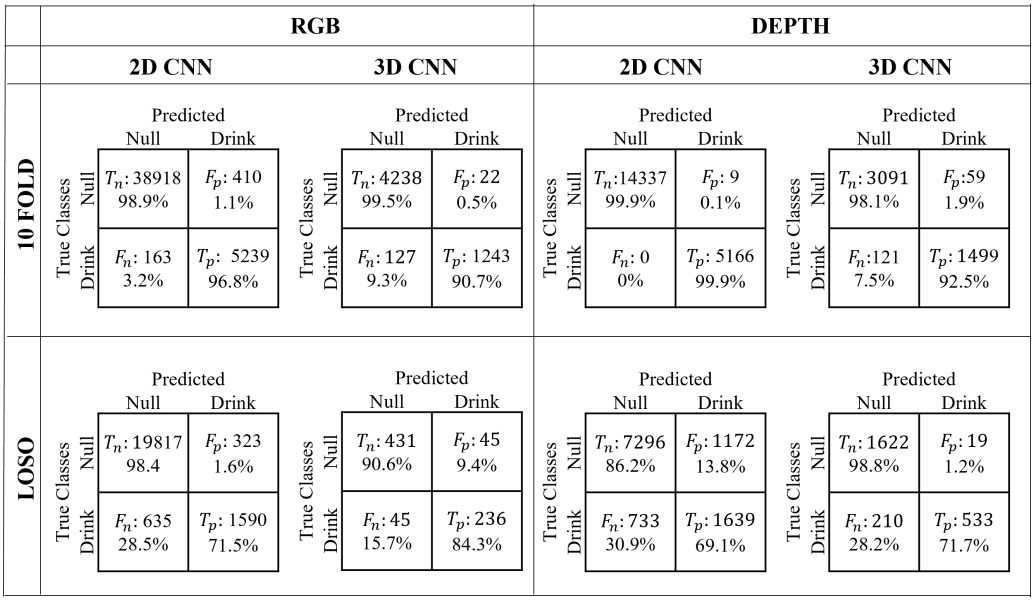

FIGURE 6. Confusion matrices for the 3D models, comparing RGB and depth inputs with 10-Fold and LOSO cross validation.

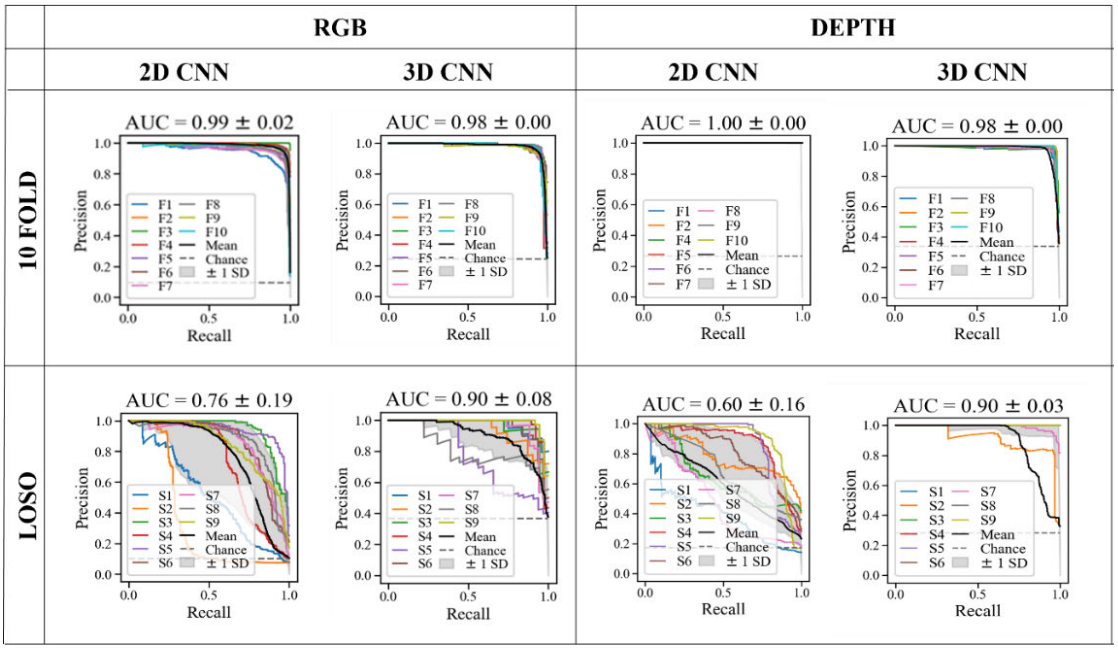

FIGURE 7. ROC curves for the best RGB and depth models.

matrices of LOSO validation show that, the 3D model trained on RGB provided closer amounts of false positives and false negatives (15.7% vs 9.5%) compared to the model trained by the depth data (28.2% vs 1.2%). In 2D models, the model trained on depth are providing more false positives and false negatives in LOSO compared to the model trained on RGB data.

The Receiver operating curves (ROC) (Figure 7) and Precision Recall curves (Figure 8) show that, in all cases, the 2D models have lower Area Under the Curve (AUC) and standard deviations compared to the 3D models. All of the 10-Fold validation models achieve a near perfect AUC.

A heat map was created using Gradient-weighted Class Activation Maps and was overlayed on top of the original images. This uses the gradients of the last convolutional layer before the output layer to localize where the model is "looking" to make the prediction (Figure 9). In most cases, the model is looking at the participant. However sometimes it is looking at other noise in the environment (Figure 9(a)) where, although it was properly classified, the model is looking at the participant, the ceiling, and the table. The example frames in Figure 9 (b) and (c) indicate that the model is properly looking into the participant's face to detect the drinking event. Future works should attempt to fully isolate the person in the frame to improve the model performance.

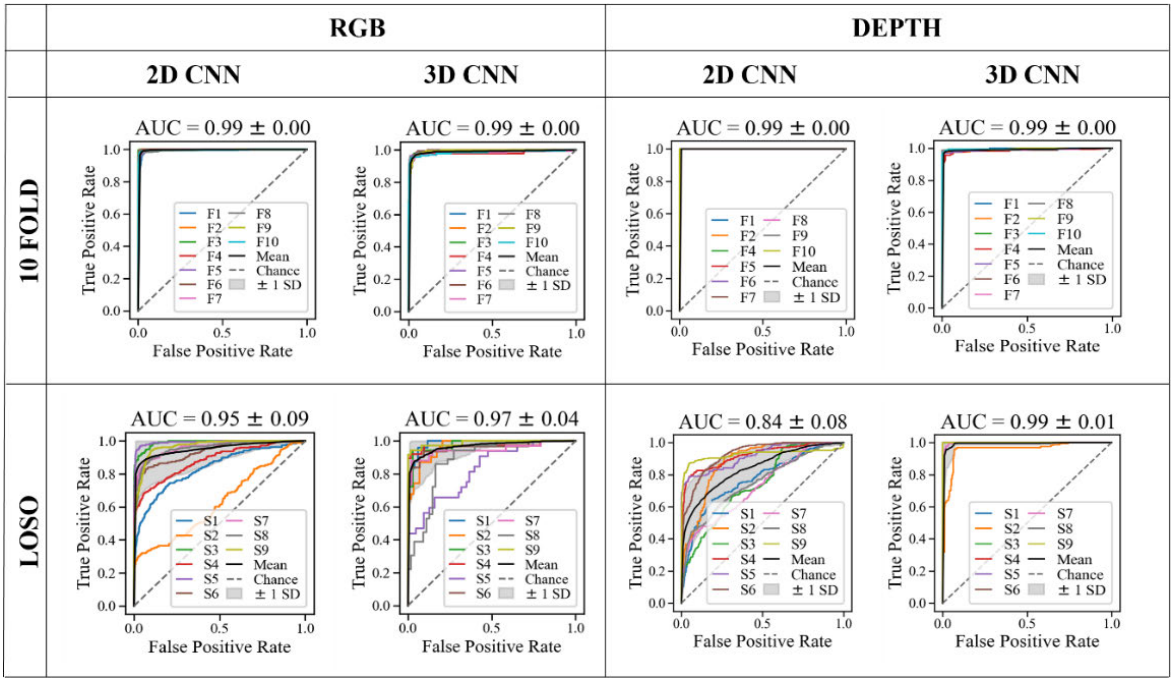

FIGURE 8. Precision recall curves for the best RGB and depth models.

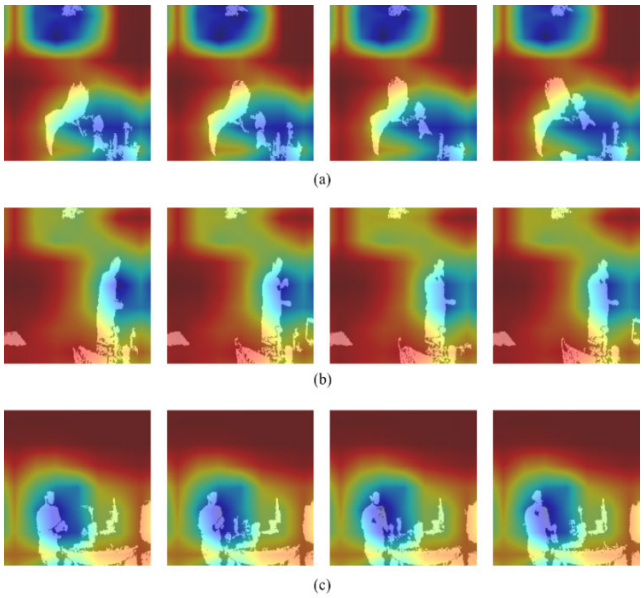

FIGURE 9. Gradient class activation maps for 3 different examples of drink events using depth data.

Despite their improved performance over the 2D models, the drawback of using 3D models is their high computational cost to train. When implementing this model, this would also increase the computational time it takes to make a prediction. The 2D model took 0.4 s to make a prediction on an individual input while the 3D model took 1.3 s. However, the 2D model needs to perform this prediction on every individual frame but the 3D model first collects 3 or 10 seconds of data to input to the model, therefore the longer computational time is not detrimental. Therefore, either the 2D or 3D models are practical for implementation in a real-world setting.

TABLE 3. Comparison with the literature using depth.

Ref.	Inputs	Method of detection	# Subs.	Acc. (%)	F1-score (%)
Tham [6]	Depth images	Dynamic Time Warping	22	89	93
Cippitelli [8]	Top Down RGB and Depth images	Self-Organized map	35 during a meal	98.3	-
Cunha [11]	Depth and RGB images	Skeleton tracking distances	3 during a meal	89	-
Rouast [15]	RGB videos	3D CNN Holdout	102	-	85.8
Proposed	RGB Videos	3D CNN LOSO	9	88.6	84.2
		3D CNN 10-fold		97.1	94.3
	Depth Videos	3D CNN LOSO		94.8	86.2
		3D CNN 10-fold		96.2	93.9

This work builds on previous works by using videos instead of images, providing a robust dataset not limited to meal time events, and compares the use of RGB to depth data separately. A comparison with the literature can be found in Table 3.

Even though depth cameras are privacy preserving, there is still the problem of occlusion - as with all vision-based approaches. The person must be visible in the frame, meaning drinks captured outside the home would not be registered. The major limitation of this study is that only the drinking action was detected and the amount of liquid intake was not measured. Real-time deployment of the final model should also be investigated further, since the 3D models would have to cache frames to create the input, before injecting them into the model. Additionally, adding other features such as overlaying human pose/joint tracking should be investigated to determine if tracking the human can improve the model's

performance. Fusing both RGB and depth into one input can also be analyzed to see if it increases the model's performance. Finally, future work should test the algorithm in real time and create a prompting system to remind the user to drink throughout the day. Since some daily liquid intake also comes from food intake, it is important to expand this algorithm to classify food intake events, particularly those from a spoon.

V. CONCLUSION

Dehydration is a common and potentially consequential issue when presented in older adults, therefore it is critical to remind older adults to drink regularly to ensure proper hydration. The first step of such a prompting system is to detect when a drink occurs and only remind the user to drink as needed. This work compares the use of depth cameras and RGB cameras to classify drinking intake events in the home environment. We have compared 2D and 3D CNN models. Overall, we found that the 3D models are more robust and have higher F1-scores than the 2D models for both depth and RGB inputs. However, for the 3D models, the depth and RGB inputs have very similar F1-scores for both 10-fold and LOSO cross validation. This shows that depth data is a viable option as it is also privacy preserving. Future work should investigate fusing the RGB and depth data or overlaying a skeleton pose tracking as data augmentation. We will also create a real-time classification and prompting system to remind the user to drink and track liquid intake events.

REFERENCES

[1] J. A. Bennett, "Dehydration: Hazards and benefits," *Geriatric Nursing*, vol. 21, no. 2, pp. 84–88, Mar. 2000.

[2] P. A. Phillips, B. J. Rolls, J. G. G. Ledingham, M. L. Forsling, J. J. Morton, M. J. Crowe, and L. Wollner, "Reduced thirst after water deprivation in healthy elderly men," *New England J. Med.*, vol. 311, no. 12, pp. 753–759, Sep. 1984, doi: 10.1056/NEJM198409203111202.

[3] R. Cohen, G. Fernie, and A. R. Fekr, "Automated fluid intake detection using RGB videos," *Sensors*, vol. 22, no. 18, p. 6747, Sep. 2022, doi: 10.3390/S22186747.

[4] R. Cohen, G. Fernie, and A. R. Fekr, "Fluid intake monitoring systems for the elderly: A review of the literature," *Nutrients*, vol. 13, no. 6, p. 2092, Jun. 2021, doi: 10.3390/NU13062092.

[5] J.-L. Chua, Y. C. Chang, M. H. Jaward, J. Parkkinen, and K.-S. Wong, "Vision-based hand grasping posture recognition in drinking activity," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, Sarawak, Malaysia, Dec. 2014, pp. 185–190, doi: 10.1109/ISPACS.2014.7024449.

[6] J. S. Tham, Y. C. Chang, and M. F. A. Fauzi, "Automatic identification of drinking activities at home using depth data from RGB-D camera," in *Proc. Int. Conf. Control, Autom. Inf. Sci.*, Gwangju, South Korea, Dec. 2014, pp. 153–158, doi: 10.1109/ICCAIS.2014.7020549.

[7] Y. C. Chang, A. R. Sheikh, J. L. Chua, and J. S. Tham, "Visual based dining activities detection in ambient assisted living," in *Proc. IEEE Int. Conf. Consum. Electron.*, Jun. 2015, pp. 494–495, doi: 10.1109/ICCE-TW.2015.7217017.

[8] E. Cippitelli, S. Gasparrini, E. Gambi, and S. Spinsante, "Unobtrusive intake actions monitoring through RGB and depth information fusion," in *Proc. IEEE 12th Int. Conf. Intell. Comput. Commun. Process. (ICCP)*, Cluj-Napoca, Romania, Sep. 2016, pp. 19–26, doi: 10.1109/ICCP.2016.7737116.

[9] E. Gambi, M. Ricciuti, and A. De Santis, "Food intake actions detection: An improved algorithm toward real-time analysis," *J. Imag.*, vol. 6, no. 3, p. 12, Mar. 2020, doi: 10.3390/JIMAGING6030012.

[10] S. Gasparrini, E. Cippitelli, E. Gambi, S. Spinsante, and F. Florez-Revuelta, "Performance analysis of self-organising neural networks tracking algorithms for intake monitoring using Kinect," in *Proc. IET Int. Conf. Technol. Act. Assist. Living (TechAAL)*, London, U.K., 2015, pp. 1–6, doi: 10.1049/IC.2015.0133.

[11] A. Cunha, L. Pádua, L. Costa, and P. Trigueiros, "Evaluation of MS Kinect for elderly meal intake monitoring," *Proc. Technol.*, vol. 16, pp. 1383–1390, Jan. 2014.

[12] L. Costa, P. Trigueiros, and A. Cunha, "Automatic meal intake monitoring using hidden Markov models," *Proc. Comput. Sci.*, vol. 100, pp. 110–117, Jan. 2016, doi: 10.1016/J.PROCS.2016.09.130.

[13] M. F. Kassim, M. N. H. Mohd, M. R. M. Tomari, N. S. Suriani, W. N. W. Zakaria, and S. Sari, "A non-invasive and non-wearable food intake monitoring system based on depth sensor," *Bull. Electr. Eng. Informat.*, vol. 9, no. 6, pp. 2342–2349, Dec. 2020, doi: 10.11591/EEI.V9I6.2256.

[14] H. M. Hondori, M. Khademi, and C. V. Lopes, "Monitoring intake gestures using sensor fusion (Microsoft Kinect and inertial sensors) for smart home tele-rehab setting," in *Proc. 1st Annu. IEEE Healthcare Innov. Conf.*, Houston, TX, USA, Nov. 2012, pp. 1–4.

[15] P. V. Rouast and M. T. P. Adam, "Learning deep representations for video-based intake gesture detection," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 6, pp. 1727–1737, Jun. 2020, doi: 10.1109/JBHI.2019.2942845.

[16] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 4724–4733, doi: 10.1109/CVPR.2017.502.

RACHEL COHEN received the B.A.Sc. degree in biomedical mechanical engineering from the University of Ottawa, in 2019, and the M.A.Sc. degree in biomedical engineering from the University of Toronto, in 2022.

She is currently working with the KITE Research Institute—Toronto Rehabilitation Institute—University Health Network. Her research interests include using artificial intelligence for medical applications, including health monitoring for older adults.

GEOFF FERNIE received the Ph.D. degree.

He is currently the Creaghan Family Chair of Prevention and Healthcare Technologies, Department of Surgery, University of Toronto, and the Toronto Rehabilitation Institute. His passion is the search for practical solutions to common problems of daily living for an aging population, people with disabilities, and their caregivers. He has a track record of taking inventions from the laboratory to market. He has commercialized 11 products and three currently in clinical trials and has helped to launch several companies. He has over 140 peer-reviewed journal articles/book chapters, 22 awarded patents, and an additional 13 patent filings. His research focus is on increasing safe mobility. He reduces falls through improved environmental design and footwear and increases safety for older drivers. He also develops and is commercializing technology to reduce the large numbers of patients, who catch infections when in hospital. He has received the P.Eng, C.Eng. He was also appointed to the Order of Canada and the Order of Ontario. He is a Fellow of the Canadian Academy of Health Sciences and a Fellow of the Canadian Academy of Engineering.

ATENA ROSHAN FEKR received the Ph.D. degree in electrical and computer engineering from McGill University, Montreal.

She is currently an Assistant Professor with the Institute of Biomedical Engineering, University of Toronto, and an Affiliate Scientist with the KITE Research Institute, University Health Network. Her primary research interests include to combine ubiquitous sensing technologies with machine learning, optimization, and signal processing techniques to solve real-world, practical problems. This includes using technology, artificial intelligence, and sensor data fusion to help elderly people and patients with chronic disease.