Robust Hand Hygiene Monitoring for Food Safety using Hand Images

Shengtai Ju[1], Amy R. Reibman[1], and Amanda J. Deering [2]
[1]School of Electrical and Computer Engineering, [2]Department of Food Science,
Purdue University, West Lafayette, Indiana, USA

Abstract

Hand hygiene is essential for food safety and food handlers. Maintaining proper hand hygiene can improve food safety and promote public welfare. However, traditional methods of evaluating hygiene during food handling process, such as visual auditing by human experts, can be costly and inefficient compared to a computer vision system. Because of the varying conditions and locations of real-world food processing sites, computer vision systems for recognizing handwashing actions can be susceptible to changes in lighting and environments. Therefore, we design a robust and generalizable video system based on ResNet50 that includes a hand extraction method and a 2-stream network for classifying handwashing actions. More specifically, our hand extraction method eliminates the background and helps the classifier focus on hand regions under changing lighting conditions and environments. The results show that our system is more robust and generalizable when evaluated on completely unseen data by achieving over 17% improvement on the overall classification accuracy.

Introduction

Food safety and the prevention of food-borne illnesses is critical to promote public welfare. To improve food safety, food handlers should follow the handwashing steps from the WHO (World Health Organization) [1] to ensure all surfaces of their hands are properly rubbed and sanitized. More specifically, the WHO handwashing steps include the following 7 rubbing actions: rub hands palm to palm, rub back of each hand, rub palm to palm with fingers interlaced, rub with back of fingers to opposing palms, rub each thumb, rub tips of fingers, and rub each wrist. Because these actions are fine-grained with high inter-class similarity, traditional methods of hand hygiene monitoring require auditing by human experts, which can be labor-intensive and inefficient. Furthermore, food handlers that have different skin tones often wash their hands at locations with various lighting conditions. Therefore, having an efficient method that is robust to domain shifts is essential for handwashing monitoring during food handling.

Compared to auditing by experts, more advanced methods using wearable devices have been developed recently. Armbands are used in [2] to classify and monitor handwashing activities. Furthermore, a handwash monitor and feedback system using a smartwatch was presented in [3]. Using wearable sensors can be much more efficient than auditing by human experts. However, wearable sensors can be intrusive and inconvenient for users, especially for food handlers who may work outdoors. Therefore, we design a video analytics system using only a camera and a laptop to assess handwashing. This is more efficient, more practical, and less intrusive compared to traditional methods and wearable devices.

Our goal is to design an accurate, robust, and generalizable action recognition system for different steps of handwashing. To build such a system, we recognize that rubbing actions are challenging to classify because of heavy occlusion and low inter-class variance. In particular, heavy occlusion occurs when two palms are overlapped or interlaced, causing hands to appear incomplete in RGB images. Moreover, changes in lighting conditions, such as indoor vs. outdoor lighting, and sunny vs. cloudy conditions make classification challenging. It is important that a handwashing monitoring system be robust and generalizable across different environments. Therefore, in this paper, we design a robust classifier for the WHO handwashing steps by incorporating both a hand extraction method and a hand pose estimator. More specifically, the hand extraction method enables us to remove the background from images and focus on the hand regions. In addition, our goal for applying hand pose estimation is to eliminate the effects of lighting changes and variations in skin tones. We also apply a 2-stream network with RGB and hand streams to further improve performance. We show that our system is robust and generalizable by evaluating the classification accuracy using 3 different datasets captured across varied environments.

The rest of this paper is structured as follows. We begin by introducing existing works regarding handwashing monitoring and recognition. Next, we discuss details about the dataset we created and the public dataset we used for evaluation. Then, we provide an overview of our system and a detailed description of each component within our system. Finally, we demonstrate our system's recognition accuracy using different evaluation datasets.

Related Work

In this section, we review existing works and systems related to handwashing recognition. In recent years, researchers have designed camera-based video systems for monitoring handwashing activities. For example, a handwashing station [4] was deployed in a school during the COVID-19 pandemic, which consisted of a camera, a UV light component, and a pressure mat. Additionally, a UV fluorescent compound was applied on hands to monitor which surfaces have been rubbed. They have shown that deploying a handwashing station is practical and useful for improving handwashing quality among children. Moreover, object tracking and a Markov decision process were applied in a real-time vision system for assisting people with dementia to wash their hands [5]. Similarly, hand detection and motion trajectory analy-

sis were used to monitor handwashing activities for older adults as potential indicators of dementia [6]. Although these systems have demonstrated the effectiveness of camera-based approaches, they do not focus on classifying the challenging fine-grained actions required in the WHO guideline.

Hand hygiene also plays an important role in health care. A handwashing monitoring system was built by first tracking hands using Particle and Kalman filters by Lacey et al. [7]. Later, a handwashing pose classifier was introduced using a multi-class SVM ensemble [8]. Next, a visioin-based system for assessing handwashing quality was designed by combining object tracking and multi-class SVM [9]. Then, RGB-D videos were utilized for classifying different handwashing actions [10]. Finally, a handwashing tutorial system was developed for deployment in hospitals [11][12][13]. Although this series of work focuses on classifying the WHO handwashing steps, they have only considered indoor scenarios in hospitals, whereas we are interested in outdoor settings for food handling. Also, the evaluation dataset used in [8] and [9] only consisted of 6 short videos. Our goal is to evaluate a system using videos captured with outdoor lighting conditions and various participants with diverse characteristics.

Moreover, with recent advancement in neural networks and computer vision, many deep-learning based handwashing systems have been developed. A two-stream network using RGB and optical flow for classifying egocentric handwashing activities was introduced by [14]. They then extended this work by adding a coarse-to-fine classification strategy and motion histogram images [15]. Additionally, a multi-view camera system for classifying hand-to-object interactions was introduced in [16]. These existing systems have focused more on hand-to-object actions and coarse-grained rubbing actions instead of fine-grained actions. Self-attention blocks were combined with a CNN feature extractor to classify challenging rubbing actions in [17]. The authors have also created and published a dataset of different handwashing actions. Another public dataset was created in a hospital setting [18]. These datasets lack outdoor lighting changes and variations in skin tones, which are important for building a robust and generalizable system.

Dataset

There are public datasets available for handwashing recognition [17][18]. However, as discussed in the related work section, existing datasets lack the outdoor lighting changes and variation in participants that we are interested in. Therefore, we recorded videos of handwashing activities using a portable sink and collected video data from a wide range of participants. All handwashing steps follow the WHO guideline as shown in Figure 1. In the following subsections, we discuss details regarding our Portable51 dataset, our Farm23 dataset, and a public Hand Wash Dataset [19] that we use for evaluation.

Portable51 Dataset

We recorded people washing their hands on a portable sink using an OAK-D camera in 1080P resolution at 30 frames per second. We invited 51 participants from the general public to wash their hands following the WHO steps. Brief instructions and demonstration of the handwashing steps were given to the participants prior to each recording. To best consider the real application settings of such a recognition system and best evaluate our system, we recorded videos under a wide range of lighting conditions, including sunny outdoor, cloudy outdoor, rainy outdoor, and indoor lighting. The participants also had different skin tones and ages. Furthermore, data collection was done across multiple days and at different locations. Figure 1 shows one sample frame for each action. Lighting changes and skin-tone variation can be seen from the sample frames.

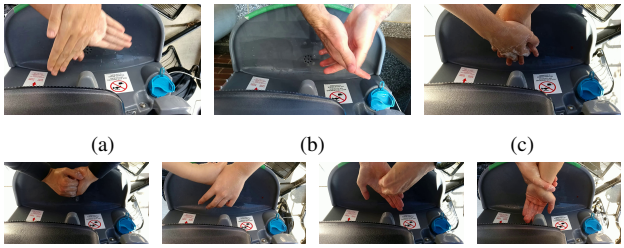

Figure 1: (a) rub palm (b) rub palm with fingers interlaced (c) rub back of hands (d) rub back of fingers (e) rub thumb (f) rub finger tips (g) rub wrist

Farm23 Dataset

We collected a second set of videos at the Purdue Student Farm, following the same recording setup as the Portable51 dataset. 12 new participants, completely different from those in Portable51, helped record videos. This dataset contains more challenging videos because the light source (the sun) is behind the participants during recording, which casts large shadows inside the field of view. 23 new handwashing videos were recorded and used for evaluating the robustness of our system.

Hand Wash Dataset

The Hand Wash Dataset [19] contains video clips of 7 rubbing actions following the same WHO guideline. Videos were recorded using different types of sinks as background. The dataset contains 300 short video clips of handwashing actions. 25 videos are available for each of these two actions: rubbing hands palm to palm and rubbing palm to palm with fingers interlaced. 50 videos are available for the other 5 actions. Figure 2 shows examples of the same action being recorded using different sinks.


Figure 2: Hand Wash Dataset Examples

Method

Overview

To reduce the effect of lighting changes and make our system more robust, we use a hand extraction method to remove the background and focus only on the hand regions which contain the fine-grained rubbing actions. Moreover, we apply a two stream network with RGB and hand streams to further leverage both high-level and detailed features. In addition, we explore applying a hand keypoint estimator and classifying hand skeletons to reduce the impact of appearance changes and domain shifts.

In this paper, we consider the handwashing-step recognition problem as an *image classification* problem because of realistic limits of computational resources in a real-time system. One of the most popular methods to incorporate motion information is to use optical flow. However, optical flow can increase computational cost significantly. To demonstrate this, we compare the run-time of three optical flow methods, Farneback [20], TV-L1 [21], and RAFT [22], against color thresholding. Farneback, TV-L1, and color thresholding are run using an Intel i7 CPU, while RAFT is run on a GPU. For 1080p images, Farneback and TV-L1 took 477ms and 6492ms, respectively. The GPU-based runtime for RAFT is 550ms on 1080p images. However, applying color thresholding on 1080p images only took 2.6ms. As can be seen from the results, color thresholding is much faster to compute compared to optical flow methods. Therefore, unlike in [14] and [23], we choose to not use optical flow in our two stream network.

Figures 3 and 5 show the block diagrams of our hand extractor and two stream network, respectively. The hand classifier is built by finetuning a ResNet50 model with hand images. In addition, the two stream network combines RGB and hand streams. Each stream is trained individually with the corresponding input images, while both streams use ResNet50. The individually trained models act as feature extractors in the two stream network. Finally, features are averaged at the end and fed into two fully connected layers for final classification.

Hand Extraction

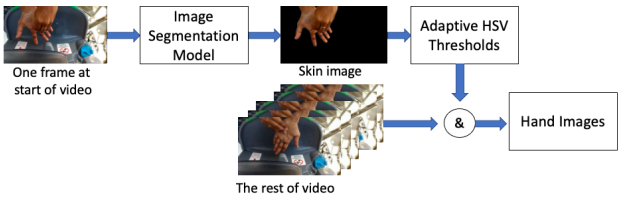

Figure 3: Hand extraction block diagram

The hand extraction method transforms RGB images into hand images by applying HSV color thresholding. For this process, we manually select one frame at the beginning of each individual long video because we need to make sure hands are present in this frame for successful hand extraction. Then, we apply a pretrained DeepLabV3[24] image segmentation model on this frame to find regions that belong to the class "human". These segmented regions contain skin areas for each individual. Next, we use these skin regions to form a set of adaptive HSV thresholds. Each channel in the HSV color space has its own threshold. More specifically, we build color histograms for each channel and define the upper bound of H channel to include 90% of the found skin regions. Additionally, the lower bounds of S and V channels are set to include 85% of skin regions. The lower bound of H channel is set to 0 and the upper bounds of S and V channels are set to 255. These percentiles are determined through empirical observation. Next, we apply the found thresholds to extract hand regions for the remaining frames of the long video. By applying this procedure to each individual, we can adjust to different lighting conditions and ensure the best quality for extracting hand images.

Figure 4 shows successful and failed examples of hand extraction. Successful hand extraction can eliminate all background including the portable sink and other irrelevant objects. On the other hand, failed hand extraction usually contains portions of the sink and background. Also, failed extraction can result in large holes in hands because of thresholding.

Two Stream Network

As can be seen in Figure 5, the two stream network utilizes the RGB stream and hand stream pretrained on our RGB images and hand images, respectively. The RGB and hand models act as feature extractors with their weights fixed and final layer removed. 2048-dimensional features from each stream are averaged and then passed through 2 fully connected layers. Our goal is to combine both high-level information from the RGB stream, and lower-level information from the hand stream to improve performance.

Hand Pose Estimation

To address the problem of appearance changes and domain shifts, we notice there are distinct hand shapes and positions during each rubbing action. For example, rubbing the back of fingers with palms interlaced requires fingers of two hands to be interlocked. Moreover, rubbing thumb requires one hand to grab the thumb of the other hand. All 7 rubbing actions require different hand positions and shapes. By considering hand poses, we believe we could eliminate the effects of varying lighting conditions and other appearance differences that impact model robustness. Thus, we explore whether the addition of hand keypoint estimation is helpful for recognizing different rubbing actions.

For hand keypoint estimation, we experiment with 3 different models [25], [26], and [27]. Through empirical observation, we choose to apply the MediaPipe Hands [25]. It estimates the 3D coordinates of 21 important hand keypoints: 4 joints for each finger and 1 point for the palm location.

Also, we combine our hand extraction method with MediaPipe. The hand keypoint estimator requires a bounding box for each hand. Therefore, we generate a bounding box for each hand from our hand images. To accomplish this, we find a large enclosing bounding box for the overall hand region inside the hand image and create two equal-sized bounding boxes for the two hands. Here, we assume that both hands are always visible in our images because of the nature of handwashing and rubbing actions. The bounding box for one hand is centered at $x + 3/8 * w$, with x representing the starting x coordinate of the larger hand-region bounding box and w representing the width of the hand-region bounding box. In addition, the bounding box for the other hand is centered at $x + 5/8 * w$. With these bounding boxes, we apply the MediaPipe hand model to estimate the coordinates of 21 hand joints. Figure 6 shows two successful examples and two failed examples of hand keypoint estimation. As can be seen, the first failed example is due to overlapping hands and heavy occlusion. The second failed example is due to only one set of fingers being visible, confusing the keypoint model to estimate fingers for both hands at the same location.

Experiments and Results

Hand Extraction and Hand Classifier

We finetune a ResNet50 model pre-trained on ImageNet [28] using our extracted hand images. Moreover, we finetune a ResNet50 model with RGB images as our baseline method, and this also serves as a feature extractor in the two stream network.

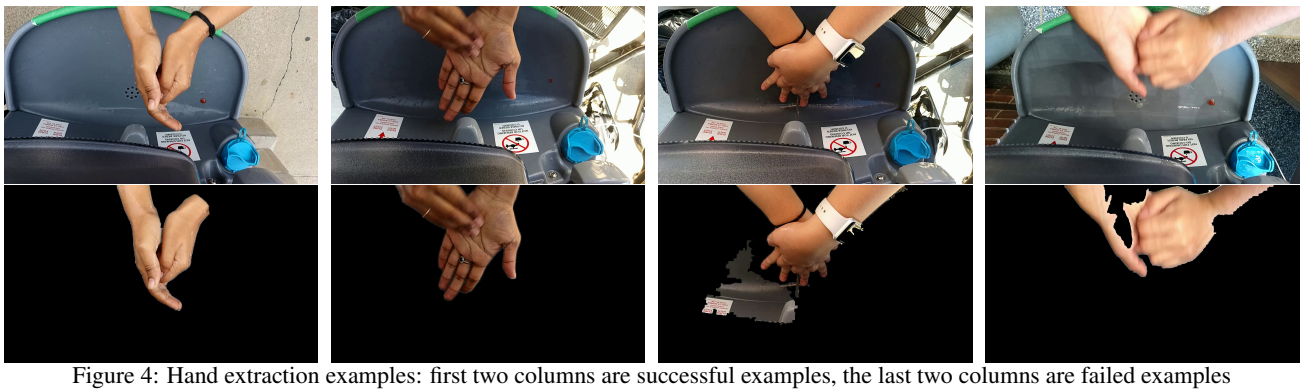

Figure 4: Hand extraction examples: first two columns are successful examples, the last two columns are failed examples

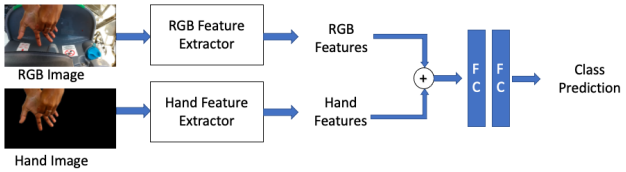

Figure 5: Two stream network with RGB and hand streams

We split the Portable51 dataset into training and testing sets using a 7:3 ratio in terms of total number of videos, i.e. 36 training videos and 15 testing videos. We split data this way instead of randomly splitting the entire collection of frames because we want to ensure there is enough lighting and skin tone variation in both training and testing. To evaluate our classifiers, we use 3 different datasets: 15 long videos of 15 different participants from Portable51 dataset, Farm23 dataset with 23 long videos of 12 different individuals, and the Hand Wash Dataset with 300 total action clips.

All models are trained for 10 total epochs with a batch size of 64. We use the SGD optimizer with a learning rate of 1e-3 and momentum of 0.9 for all model training. Training and testing is done using a single NVIDIA TITAN GPU. In the tables, the best performance per row is highlighted in bold.

Hand Pose Estimation

With 21 hand joints estimated, we experiment with 4 different feature representations. The first representation is the normalized 3D XYZ coordinates of the 21 hand joints. The second representation is a simple hand skeleton with 5 vectors, from the palm point to each fingertip. The third representation is a pre-defined full hand skeleton with 21 vectors, as shown in Figure 7. The last representation is the combination of normalized 3D coordinates and full 21-vector skeleton.

For classification, we choose to compare results using 3 different classifiers: a multi-class SVM, a Random Forest classifier, and a 2-layer MLP. The training and testing data split is the same as in the experiments for hand images. We evaluate by training on Portable51 data and testing on both Portable51 and the Hand Wash Dataset.

Hand Classifier Results

We evaluate our hand classifier and two stream network using top-1 classification accuracy. As can be seen from Table 1, the two stream network achieves the highest overall classification accuracy of 72.4%. The hand classifier outperforms the baseline RGB model by 1.9% in terms of overall accuracy. Moreover, the hand classifier outperforms the RGB model for 5 out of 7 rubbing actions. The two stream network outperforms the hand classifier by another 2.4% and the RGB classifier by 4.3% in overall accuracy. In addition, it obtains higher individual-action accuracy for 5 out of 7 actions compared to the hand classifier. When compared to the RGB model, the two stream network is better for 6 out of 7 rubbing actions, except for rubbing back of fingers.

Action (# of images)	RGB	Hand	Two Stream
Rub back (2811)	**83.6%**	81.0%	83.0%
Rub back fingers (2784)	53.7%	**63.1%**	60.5%
Rub palm (2502)	79.6%	77.0%	**83.6%**
Rub palm fingers interlaced (1687)	47.4%	49.4%	**52.0%**
Rub thumb (2475)	78.1%	74.9%	**79.6%**
Rub tips (2481)	48.6%	52.3%	**56.5%**
Rub wrist (2905)	77.4%	83.0%	**83.4%**
Overall (17645)	68.1%	70.0%	**72.4%**

Table 1: Results on Portable51

Action (# of images)	RGB	Hand	Two Stream
Rub back (19490)	17.3%	**64.7%**	49.2%
Rub back fingers (20743)	52.5%	50.3%	**69.3%**
Rub palm (11735)	9.5%	58.1%	**64.8%**
Rub palm fingers interlaced (9755)	7.1%	31.0%	**33.8%**
Rub thumb (19946)	2.3%	**51.8%**	16.5%
Rub tips (17626)	**34.6%**	12.7%	24.2%
Rub wrist (16608)	**97.0%**	81.5%	96.4%
Overall (115903)	33.4%	**50.9%**	50.4%

Table 2: Results on Hand Wash Dataset

Table 2 shows the evaluation results using the Hand Wash Dataset [19]. Because this dataset contains different types of sinks that are significantly different from our portable sink, the RGB classifier suffers from domain shifts in the background region. The hand classifier achieves the highest overall classification accuracy, slightly outperforming the two stream network by 0.5%. However, both the hand classifier and two stream network outperform the RGB model by over 17% overall. The addition of hand

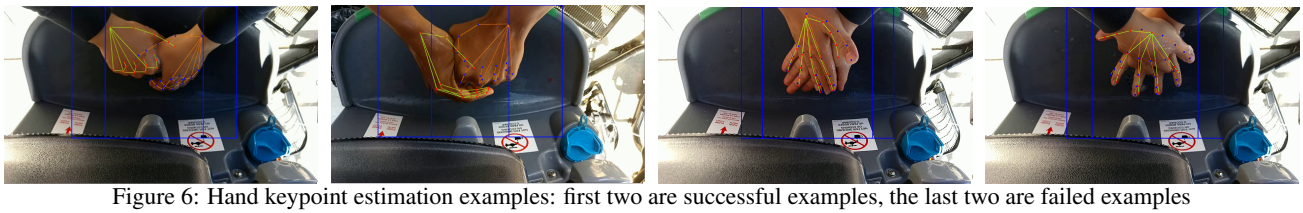

Figure 6: Hand keypoint estimation examples: first two are successful examples, the last two are failed examples

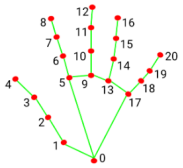

Figure 7: Hand Skeleton

images makes the combination of hand classifier and two stream network outperform the RGB model in 5 out of 7 actions. The most significant changes occur for the actions of rubbing palm and rubbing palm with fingers interlaced. For rubbing palm, the classification accuracy increased from 9.5% to over 58.1% for hand classifier and two stream network. For rubbing palm with fingers interlaced, the accuracy increased from 7.1% to over 31.0% when using hand images. These results show that when evaluated on completely unseen data with significant domain changes, our hand extraction method is able to improve model's robustness and the ability to generalize.

Action (# of images)	RGB	Hand	Two Stream
Rub back (1334)	43.9%	**54.7%**	53.1%
Rub back fingers (1545)	38.8%	39.8%	**45.4%**
Rub palm (1221)	19.2%	**39.3%**	36.4%
Rub palm fingers interlaced (1109)	30.4%	27.2%	**33.1%**
Rub thumb (1487)	43.4%	**69.4%**	65.3%
Rub tips (1830)	**58.5%**	41.5%	53.9%
Rub wrist (1482)	80.2%	75.2%	**83.9%**
Overall (10008)	46.6%	50.3%	**54.2%**

Table 3: Results on Farm23

Table 3 shows the results when evaluating on the Farm23 dataset. In this case, the two stream network performs the best with an overall accuracy of 54.2%, outperforming the hand classifier by 3.9% and the RGB classifier by 7.6%. The two classifiers with the hand extraction method outperform the baseline classifier in 6 out of 7 actions. The most significant increase in performance occurs for the rubbing thumb action. Classification accuracy increased from 43.4% to over 65.3% for both hand and two stream classifiers, which is an increase of over 21%. These results further demonstrate the added robustness and generalizability of using hand images.

Hand Pose Estimation Results

We report the overall classification accuracy for Portable51 in Table 4 and Hand Wash Dataset in Table 5 when we incorporate the hand pose. As can be seen, hand pose estimation failed to achieve satisfactory results even on the Portable51 data. One of the main reasons this method failed is the inconsistency in hand skeleton quality. Under heavy occlusion, the estimated hand keypoints are usually inaccurate, which lead to an inaccurate hand skeleton and hand pose. Another cause of unsatisfactory performance is high intra-class variance, which means different individuals perform the same rubbing action in different poses and hand orientations. Although this method attenuates the environmental and appearance effects, its performance is limited for our application currently.

Classifier	XYZ only	Skeleton5	Skeleton21	XYZ + Skeleton21
SVM	50.5%	42.3%	51.2%	**51.5%**
Random Forest	47.7%	47.1%	50.4%	**53.7%**
MLP	47.3%	45.0%	**50.2%**	50.1%

Table 4: Results on Portable51 using Hand Keypoints

Classifier	XYZ only	Skeleton5	Skeleton21	XYZ + Skeleton21
SVM	41.5%	38.7%	40.4%	**41.8%**
Random Forest	31.4%	38.6%	**43.5%**	42.0%
MLP	36.8%	42.2%	**43.0%**	40.7%

Table 5: Results on Hand Wash Dataset using Hand Keypoints

Conclusion

In this paper, we describe the design and implementation of a classifier for monitoring handwashing actions following the WHO steps. We applied a hand extraction method, a two stream network, and a hand pose estimator to build robust and generalizable classifiers for challenging fine-grained rubbing actions. More specifically, the hand extraction method combines image segmentation and HSV color thresholding to generate hand images. By removing the background and focusing on the hand regions, the hand classifier improves significantly on data from different domains. By using a two-stream network with RGB and hand streams, we are able to further improve the model's performance in 2 of 3 evaluation datasets we have used. The baseline RGB classifier, however, is less robust and more vulnerable to environmental changes. Finally, we studied the application of hand pose estimation to reduce the impact of appearance changes and domain shifts. However, our experiments and results do not show improved robustness and generalizability when applying hand pose estimation. For future research, we will study how motion information can be efficiently incorporated into a handwashing system to further improve its performance without increasing computational cost significantly.

References

[1] F. G. P. S. Challenge, "WHO guidelines on hand hygiene in health care: a summary," *Geneva: World Health Organization*, vol. 119, no. 14, pp. 1977–2016, 2009.

[2] C. Wang, Z. Sarsenbayeva, X. Chen, T. Dingler, J. Goncalves, V. Kostakos *et al.*, "Accurate measurement of handwash quality using sensor armbands: Instrument validation study," *JMIR mHealth and uHealth*, vol. 8, no. 3, p. e17001, 2020.

[3] S. Samyoun, S. S. Shubha, M. A. S. Mondol, and J. A. Stankovic, "iwash: A smartwatch handwashing quality assessment and reminder system with real-time feedback in the context of infectious disease," *Smart Health*, vol. 19, p. 100171, 2021.

[4] J. Herbert, C. Horsham, H. Ford, A. Wall, E. Hacker *et al.*, "Deployment of a smart handwashing station in a school setting during the COVID-19 pandemic: Field study," *JMIR Public Health and Surveillance*, vol. 6, no. 4, p. e22305, 2020.

[5] J. Hoey, P. Poupart, A. von Bertoldi, T. Craig, C. Boutilier, and A. Mihailidis, "Automated handwashing assistance for persons with dementia using video and a partially observable Markov decision process," *Computer Vision and Image Understanding*, vol. 114, no. 5, pp. 503–519, 2010.

[6] A. Ashraf and B. Taati, "Automated video analysis of handwashing behavior as a potential marker of cognitive health in older adults," *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 2, pp. 682–690, 2015.

[7] I. Parra, D. Fernandez, M. Sotelo, M. Marron, M. Gavilan, and G. Lacey, "Tracking using Particle and Kalman filters in hand washing quality assessment system," in *2007 IEEE International Symposium on Intelligent Signal Processing*. IEEE, 2007, pp. 1–6.

[8] D. Llorca, F. Vilarino, Z. Zhou, and G. Lacey, "A multi-class SVM classifier ensemble for automatic hand washing quality assessment," in *BMVC Proc. Brit Mach Vision Conference, Warwick, UK*. sn, 2007, pp. 213–223.

[9] D. F. Llorca, I. Parra, M. Á. Sotelo, and G. Lacey, "A vision-based system for automatic hand washing quality assessment," *Machine Vision and Applications*, vol. 22, no. 2, pp. 219–234, 2011.

[10] B. Xia, R. Dahyot, J. Ruttle, D. Caulfield, and G. Lacey, "Hand hygiene poses recognition with RGB-D videos," in *Proceedings of the 17th Irish Machine Vision and Image Processing Conference*. Irish Pattern Recognition & Classification Society, 2015, pp. 43–50.

[11] A. J. Stewardson, A. Iten, V. Camus, A. Gayet-Ageron, D. Caulfield, G. Lacey, and D. Pittet, "Efficacy of a new educational tool to improve handrubbing technique amongst healthcare workers: A controlled, before-after study," *PLoS One*, vol. 9, no. 9, p. e105866, 2014.

[12] G. Lacey, M. Showstark, and J. Van Rhee, "Training to proficiency in the WHO hand hygiene technique," *Journal of Medical Education and Curricular Development*, vol. 6, p. 2382120519867681, 2019.

[13] G. Thirkell, J. Chambers, W. Gilbart, K. Thornhill, J. Arbogast, and G. Lacey, "Pilot study of digital tools to support multimodal hand hygiene in a clinical setting," *American Journal of Infection Control*, vol. 46, no. 3, pp. 261–265, 2018.

[14] C. Zhong, A. R. Reibman, H. M. Cordoba, and A. J. Deering, "Hand-hygiene activity recognition in egocentric video," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing*. IEEE, 2019, pp. 1–6.

[15] C. Zhong, A. R. Reibman, H. A. Mina, and A. J. Deering, "Multi-view hand-hygiene recognition for food safety," *Journal of Imaging*, vol. 6, no. 11, p. 120, 2020.

[16] ——, "Designing a computer-vision application: A case study for hand-hygiene assessment in an open-room environment," *Journal of Imaging*, vol. 7, no. 9, p. 170, 2021.

[17] T. Xie, J. Tian, and L. Ma, "A vision-based hand hygiene monitoring approach using self-attention convolutional neural network," *Biomedical Signal Processing and Control*, vol. 76, p. 103651, 2022.

[18] M. Lulla, A. Rutkovskis, A. Slavinska, A. Vilde, A. Gromova, M. Ivanovs, A. Skadins, R. Kadikis, and A. Elsts, "Hand-washing video dataset annotated according to the World Health Organization's hand-washing guidelines," *Data*, vol. 6, no. 4, p. 38, 2021.

[19] A. Nagaraj, M. Sood, C. Sureka, and G. Srinivasa, "Sample: Hand wash dataset," (accessed September 7, 2021). [Online]. Available: https://www.kaggle.com/datasets/realtimear/hand-wash-dataset

[20] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian Conference on Image Analysis*. Springer, 2003, pp. 363–370.

[21] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime TV-L1 optical flow," in *Joint Pattern Recognition Symposium*. Springer, 2007, pp. 214–223.

[22] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *European Conference on Computer Vision*. Springer, 2020, pp. 402–419.

[23] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[24] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.

[25] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "MediaPipe Hands: On-device real-time hand tracking," *arXiv preprint arXiv:2006.10214*, 2020.

[26] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7291–7299.

[27] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee, "InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image," in *European Conference on Computer Vision*. Springer, 2020, pp. 548–564.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255.

Author Biography

Shengtai Ju is a PhD student of Electrical and Computer Engineering at Purdue University. Shengtai received his B.S.E.E in Electrical and Computer Engineering from Purdue University in 2019.

Amy R. Reibman is the Elmore Professor of the School of Electrical and Computer Engineering at Purdue University.

Dr. Amanda Deering is an Associate Professor and Fresh Produce Food Safety Specialist in the Department of Food Science at Purdue University. Amanda is internationally known for her expertise in produce food safety and has widespread impact on safe growing and postharvest handling practices for fruits and vegetables that target growers, manufacturers, state agencies, and similar groups. She teaches both Good Agricultural Practices (GAPs) and Good Manufacturing Practices (GMPs) to fresh produce growers and food processors.