WHO-Hand Hygiene Gesture Classification System

Rashmi Bakshi

Abstract — The recent ongoing coronavirus pandemic has highlighted the importance of hand hygiene practices in our daily lives, with governments and health authorities around the world promoting good hand hygiene practices. More than 1 million cases of hospital-acquired infections are reported in Europe annually. Hand hygiene compliance may reduce the risk of cross-transmission thereby reducing the number of infections as well as health-care expenditures. In this paper, WHO hand hygiene gestures were recorded and analyzed with the construction of an aluminum frame, placed at the laboratory’s sink. The hand hygiene gestures were recorded for thirty participants after conducting a training session about hand hygiene demonstration. The video recordings were converted into image files and were organized in six different hand hygiene classes. The ResNet-50 framework was selected for the classification of multi-class hand hygiene stages. The model was trained with the first set of classes (Fingers Interlaced, P2PFingers Interlaced, and Rotational Rub) for 25 epochs. An accuracy of ~44% was achieved for the first set of experiments with loss score >1.5 in validation set. The training steps for the second set of classes (Rub hands palm to palm, Fingers Interlocked, Thumb Rub) were 50 epochs. An accuracy of ~72% was achieved for the second set with the loss score < 0.8 for the validation set. In this work, a preliminary analysis of robust hand hygiene dataset with transfer learning was carried out with a future aim of deploying a hand hygiene prediction system for healthcare workers in real-time.

Index Terms—Hand hygiene, hand washing, computer vision, deep learning, transfer learning, image recognition.

I. INTRODUCTION

Hospital Acquired Infections (HAIs) have a significant impact on quality of life and result in an increase in health care expenditure. According to the European Centre for Disease Prevention and Control (ECDC), 2.5 million cases of HAIs occur in European Union and European Economic Area (EU/EAA) each year, corresponding to 2.5 million DALYs (Disability Adjusted Life Year) which is a measure of the number of years lost due to ill health, disability or an early death [1]. MRSA-Methicillin Resistant Staphylococcus Aureus is a common bacteria associated with the spread of HAIs [2].

One of the methods to prevent the cross transmission of these microorganisms is the implementation of well-structured hand hygiene practices. The World Health Organization (WHO) has provided guidelines about hand washing procedures for health care workers [3]. Best hand hygiene practices have been proven to reduce the rate of MRSA infections in a health care setting [4].

One of the challenges in dynamic healthcare environments is to ensure compliance with these hand hygiene guidelines and to evaluate the quality of hand washing. This is often done through auditing involving human observation. The hand washing process, however, is well structured and has particular dynamic hand gestures associated with each hand washing stage.

The assessment of the process may therefore be suited to automation. Existing technology includes the use of electronic counters and RFID badges to measure the soap usage and location based reminder systems to alert the workers about washing hands [5]-[7]. These systems have shown to improve the frequency of hand washing but they do not assess if the process of handwashing is compliant with the guidelines.

One of the potential approaches is to use imaging techniques to detect fine hand movements and identify user gestures, provide feedback to the user or a central management system, with the overall goal being an automated tool that can ensure compliance with the hand washing guidelines. In advance of developing these systems, however, preliminary analysis on the hand washing process and a structured methodology was required.

The aim of this paper is to develop a classification system for WHO-hand hygiene gestures that can be deployed in a healthcare setting and assess the hand hygiene compliance in real time. Deep learning solutions were adopted for this study as they have shown promising results in other applications such as text recognition [8], sound prediction [9], and image annotation [10]. The remainder of this paper is organized as follows: Section II explains the related work. Section III provides experiment setup; section IV discusses the hand-hygiene dataset collection in detail. Section V explains ResNet architecture used for processing the hand hygiene video recordings. Section VI and VII presents the experimental results and conclusion.

II. RELATED WORK

This section discusses the past research in the field of gesture recognition with various data acquisition methods such as 3D sensors/wearable devices, and cameras. The necessity of having a robust hand hygiene data was identified at an early stage and an experiment setup was built for recording hand hygiene data. The objective behind the video data collection was to explore the paradigms of transfer learning (transfer the knowledge from one task to another) for the purpose of hand hygiene stage classification.

A. WHO Hand Hygiene Stages

There are structured and distinct guidelines for washing
hands as provided by health care authorities, WHO. These guidelines consist of eleven sequenced stages. WHO hand hygiene stages are shown in Fig. 1 with stages 2-7 which directly involves hand washing, and are the focus of this paper, with the other stages related to turning on water, drying hands [3]. Stages 2-7 were carefully analyzed and recorded with the help of 30 participants in this study. For the purpose of consistency, the names/labels of the stages are renamed and are listed in Table I.

Fig. 1. WHO Hand Hygiene Guidelines [3].

B. Gesture Recognition with 3D Sensors

Previous work with the use of various commercial gesture tracking devices such as Leap Motion Controller, Microsoft Kinect has been carried out for the purpose of detecting hand gestures in real-time. Fabio et al. [11] has extracted the hand region from the depth map and segmented it into palm and finger samples. Distance features between the palm center and the fingertips were calculated to recognize various counting gestures. Lin Shao [12] has extracted the fingertip position and palm center with the help of the Leap Motion Controller for tracking hand gestures. Distance between fingertip and palm center and the distance between two fingers adjacent to each other was calculated. Velocity was detected to differentiate between static and dynamic gestures. In our previous work, hand features were identified for various hand-hygiene stages and the Leap Motion Controller was utilized to differentiate between stationary and moving hand by extracting the palm velocity vector [13]. Marin et al. [14] have used the Leap Motion Controller and the Microsoft Kinect jointly to extract the hand features such as fingertip position, hand center, and hand curvature for recognizing the American Sign Language gestures. Jin et al. [15] have used multiple Leap Motion sensors to develop a hand tracking system for object manipulation where the sum of distal interphalangeal, proximal interphalangeal and metacarpophalangeal angles was taken into account. Strength of grasping and pinching were the object manipulation tasks that were recorded. 3D gesture trackers have shown promising result in the past in context of gesture recognition. However, they were not explored previously for tracking hand hygiene stages. Rashmi et al. [13] has detected basic hand hygiene stage, “Rub hands palm to palm” by extracting unique hand features such as palm orientation, palm curvature and distance between the two palms with the use of threshold values.

C. Gesture Recognition with Images

The field of computer vision and image processing is greatly explored in the context of gesture recognition in the past. Vision based systems and applications were built for gesture tracking and detection. Khan [16] has used color segmentation and template matching technique to detect American Sign Language gestures. Jophin et al. [17] have developed a real time finger tracking application by identifying the red color caps on the fingers using color segmentation technique in image processing. In our previous work, color techniques based on YCBCR model were explored for hand segmentation for various hand hygiene poses [13]. Azad et al. [18] has extracted the hand gesture by image segmentation and morphological operation for American Sign Language gestures. Cross correlation coefficient was applied on the gesture to recognize it with an overall 98.8 % accuracy. Chowdary et al. [19] has detected the number of circles to determine the finger count in real-time, where in the scanning algorithm is independent of the size and rotation of the hand. Liorca et al. [20] has classified the hand hygiene poses using a traditional machine learning approach with a complex skin-color detection, particle-filtering model for tracking hands. In our previous work, automated hand tracking for hand hygiene stages was carried out by extracting the image contours and centroid (cx, cy) derived from the image moments [13]. The limitation of this work was the inability to classify and predict various hand hygiene stages with centroid extraction. Center of the mass was an interesting feature for real-time hand tracking. However, it provided limited information for distinguishing one hand hygiene stage from another. Therefore, deep learning/transfer learning technology was adapted for further processing of hand hygiene stages.

D. Gesture Recognition with Deep Learning

Deep learning is an emerging approach and has been widely applied in traditional artificial intelligence domains such as semantic parsing, transfer learning, computer vision, natural language processing and more [21]. Over the years, deep learning has gained increasing attention due to the significant low cost of computing hardware and access to high processing power (eg-GPU units) [21]. Conventional machine learning techniques were limited in their ability to process data in its natural form. For decades, constructing a machine learning system required domain expertise and fine engineering skills to design a feature extractor that can transform the raw data (example: pixel values of an image) in-to a feature vector, which is passed to a classifier for pattern recognition [22]. Deep learning models learn features directly from the data without the need for building a feature extractor. Researchers have utilized deep learning/transfer learning for monitoring hand gestures. Yeung et al. [23] has presented a vision-based system for hand hygiene monitoring based on deep learning. The system uses depth sensor instead of the full video data to preserve privacy. The data is
classified using a convolutional neural network (CNN). Li et al. [24] has conducted an investigation of CNN for gesture recognition and achieved a high accuracy, demonstrating that this approach is suitable for the task. Yamamoto et al. [25] have used vision-based systems and a CNN for a hand-wash quality estimation. They compare the quality score of the automated system with a ground-truth data obtained using a substance fluorescent under ultraviolet light. Their results show that the CNN is able to classify the washing quality with high accuracy. Ivanovs et al. [26] trains the neural network on labelled hand washing dataset captured in a health care setting, applies pre trained neural network models such as MobileNetV2 and Xception with >64% accuracy in recognizing hand washing gestures. In this work, a classification model based on ResNet architecture is presented for hand gesture recognition in the hand hygiene application. The feature extraction for hand gestures with Python/OpenCV is discussed in the previous work [27].

E. Transfer Learning- a Subset of Machine Learning

Deep learning models essentially require thousands of data samples and heavy computational resources such as GPU for accurate classification and prediction analysis. However, there is a branch of machine learning, popularly known as “transfer learning” that does not necessarily require large amounts of data for evaluation. Transfer learning is a machine learning technique wherein a model developed for one task is reused for the second related task. It refers to the situation where “finding” of one setting is exploited to improve the optimization in another setting [28]. Transfer learning is usually applied to the new dataset, which is generally smaller than the original data set used to train the pre-trained model. Hussain et al. [28] has applied transfer learning to train caltech face data set with 450 face images with pre-trained model on ImageNet data set. Increasing the number of training steps (epochs) has increased the classification accuracy but it has also increased the training time as well. Computational power and time consumption were the main limitations within the study that can be overcome with the use of advanced GPUs in future work. Keras API [29] provides the most common workflow of transfer learning in context of deep learning. They are:

1) Take layers from a previously trained model
2) Freeze them to avoid destroying any of the information that they contain during the future training rounds.
3) Add some new, trainable layers on top of the frozen layers. They will learn to turn the old features into predictions on a new dataset.
4) Train the new layers on your dataset.

In this paper, these steps are the basis of constructing a model where the head of the model is replaced with a new set of fully connected layers with random initializations.

III. EXPERIMENT SETUP

An experiment setup, imitating the real world health care setting was required for the onset of this project. In order to establish the setup, a science laboratory with incorporating a sink, water tap and utilities such as soap dispenser and hand towel was selected. The assessment of the process may therefore be expanded to deploy a ‘hand hygiene prediction system’ in a healthcare setting. The system will predict the hand hygiene movements in real time and provide feedback to the user if the hand gesture is in accordance with WHO hand hygiene guidelines.

An aluminum rig was constructed and placed around the laboratory’s sink in order to incorporate a camera device for recording the hand hygiene gestures. It was built with the dimensions of 1×0.8×0.8 m (L×W×H) to ensure enough space to fit on a sink and accommodate utilities such as soap dispenser and a hand towel.

The size of the rig was determined by various factors:

- The viewing range for the ELP USB camera-2.1mm lens wide angle; 60 fps
- Controlled background exposure to avoid the skin colored objects to be miss-classified as an actual skin when processing with color based models [30]. Green and white sheets were used to minimize the background information (Fig. 2, 3).
- Maintain the privacy of the user by focusing only on the hand gestures. The height of the frame was reduced from 1 m to 0.8 m in order to avoid the appearance of body organs in the frame other than the hands.
- Enough space to fit on a sink and accommodate utilities such as soap dispenser, hand-towel.
- Lightweight yet sturdy so it can be relocated if necessary.

III. EXPERIMENT SETUP

Fig. 2. Hand Hygiene Data Collection Setup ( Hands in frame).

Fig. 3. Background information is hidden with the help of green sheets; WHO poster is displayed so the participant does not need to memorize the hand hygiene gestures.

<table>
<thead>
<tr>
<th>TABLE I. CLASS LABELS FOR HAND HYGIENE STAGES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Stage</td>
</tr>
<tr>
<td>Rub hands Palm to Palm</td>
</tr>
<tr>
<td>Right palm over left dorsum with interlaced fingers</td>
</tr>
<tr>
<td>Palm to Palm with fingers interlaced</td>
</tr>
<tr>
<td>Backs of fingers with fingers interlocked</td>
</tr>
<tr>
<td>Rotational rubbing of thumb</td>
</tr>
</tbody>
</table>
The data collection process is demonstrated with the help of experiment setup images in Fig. 2 and Fig. 3. Table I lists the class labels selected for each hand hygiene stage as per WHO guidelines.

### IV. HAND HYGIENE DATASET COLLECTION

Thirty volunteers have participated in this study. All the subjects were students ranging from undergraduate to postgraduate studies at Technological University, Dublin, Ireland. The required training and the demonstration about WHO hand hygiene guidelines was provided to the participants before the onset of an experiment. All participants have given their informed consent for inclusion before their participation. The user privacy was ensured by capturing hand and arm movements in isolation. The video length for the hand hygiene activity was recorded for 25-30 seconds. Every hand hygiene step was followed by a pause where in the user was instructed to move their hands away from the camera. Video format for this data set is MP4 file with a size of range 40-60 MB and a frame rate of 29.84 frames/s. All of the six hand washing movements were recorded in one video for each participant. One sample frame for all the stages extracted from the video files is shown in Fig. 4. For further processing, the video files were segmented into image files and six distinct classes were prepared. Table 2 lists the total number of image files for each class along with the label of the class corresponding to the original WHO hand hygiene stage.

![Sample frames collected for hand hygiene video recordings.](image)

**Fig. 4. A Sample of frames collected for hand hygiene video recordings.**

<table>
<thead>
<tr>
<th>Class Label</th>
<th>Number of Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rub hands Palm to Palm</td>
<td>2,042</td>
</tr>
<tr>
<td>Fingers Interlocked</td>
<td>1,839</td>
</tr>
<tr>
<td>Thumb Rub</td>
<td>2,019</td>
</tr>
<tr>
<td>P2PFingers Interlaced</td>
<td>2,149</td>
</tr>
<tr>
<td>Fingers Interlaced</td>
<td>2,043</td>
</tr>
<tr>
<td>Rotational Rub</td>
<td>1,834</td>
</tr>
</tbody>
</table>

The data in these classes are evenly distributed in order to avoid the bias during the training of the model.

### V. RESNET-50 ARCHITECTURE

ResNet architecture has gained immense popularity after winning the ILSVRC 2015 classification competition and COCO competitions. It is evaluated on ImageNet 2012 classification dataset that consists of 1000 classes and 1.28 million training images [31].

The ResNet architecture is based on the residual network. As the depth of the network increases, accuracy starts to become saturated and then degradation problem is exposed. The network begins to converge. Adding more layers leads to a higher training error as reported in [32]. The degradation of the network is not a result of an overfitting but the initialization of the network, optimization function or gradients. The problem of degradation is addressed by a deep residual learning framework in which stacked layers fit a residual mapping instead of the original mapping. The hypothesis is that it is easier to optimize the residual mapping than to optimize the original, unreferenced mapping. Fig 5 is the building block of the residual learning where $F(x) + x$ can be realized by identity short cut connections and their output is added to the output of the stacked layers. The basis is 3X3 filters of VGGNet model as a plain network. The short cut connections are added to transform the plain network into the residual network [33].

Resnet50 is 50 layer network, a variant based on original 34 layer Resnet where each 2 layer block in 34 layer network is replaced by 3-layer bottleneck blocks to improve the accuracy and reduce the training time; ImageNet database [33].

It consists of five convolutional layers, conv1, conv2_x, conv3_x, conv4_x, conv5_x. Once the input image is loaded, it is passed through a convolutional layer with 64 filters and a kernel size of 7x7 (conv1 layer) followed by max pooling layers. From conv2_x, the layers are grouped in pairs until the fifth convolutional layer because of the nature of the residual networks. Average pooling is applied at the fully connected layer, followed by the softmax for classification [34].

![Building block for residual learning](image)

**Fig. 5. A building block for residual learning [34].**

In this work, a pre-trained CNN model, ResNet50 [50 layers deep] on ImageNet weights is applied. The network is pre-trained on more than one million images of ImageNet dataset and can classify images into 1000 object categories. Weights="imagenet";include top= False is selected as the head of the model was replaced with a new set of fully connected layers with random initializations. All layers below the head are frozen so that their weights cannot be updated. ‘layer.trainable = False’. The model implementation is adapted from [35] where the author has implemented a multi classification system for ‘sports’ related video recordings and has achieved an accuracy higher than 90%.

### VI. EXPERIMENTAL RESULTS

The first set of experiment, Set 1 with images in multi classes- Fingers Interlaced, P2PFingersInterlaced and Rotational Rub were passed as an input to Resnet-50 network for 25 epochs. The training time consumed was ~15 hrs. Fig. 6 is the loss-accuracy curve achieved after training the model. Cross-entropy loss is the default loss function for a multi-class...
classification problem. It can be specified in Keras library as ‘categorical_crossentropy’ when compiling the model. It can be seen that the network has converged with reasonable high loss for training and validation set. The accuracy <50% results in the incorrect class predictions for hand hygiene video recordings. For the second set of experiment, Set 2, the images in multi class- Rub Palm to Palm, Thumb Rub and Fingers Interlocked were passed as an input to ResNet-50 network for 50 epochs/ training steps. The training time taken was ~ 52 hrs. Fig. 6, Fig. 7 is the loss-accuracy curve for training and validation set for Set 1 and 2 respectively and it can be observed that Set 2 loss is reasonable low in comparison to Set 1. Increase in the number of epochs/ training steps resulted in decrease in loss and higher accuracy >70%. The evidence can be seen in Fig. 8 with correct class predictions for hand hygiene video recordings [top] and incorrect predictions [bottom] for Set 1 with 25 epochs. The complete dataset with python code and results can be accessed from the online repository that is created for this project. The link is shared in the supplementary materials.

The support is the number of occurrences of each class in y_true (Target values) [36].

### Table III: Classification Report for Set 1

<table>
<thead>
<tr>
<th>Class Label</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fingers Interlocked</td>
<td>0.95</td>
<td>0.26</td>
<td>0.41</td>
<td>511</td>
</tr>
<tr>
<td>P2PFingersInterlaced</td>
<td>0.39</td>
<td>0.99</td>
<td>0.56</td>
<td>537</td>
</tr>
<tr>
<td>Rotational Rub</td>
<td>1.00</td>
<td>0.00</td>
<td>0.01</td>
<td>459</td>
</tr>
<tr>
<td>Micro average</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
<td>1507</td>
</tr>
<tr>
<td>Macro average</td>
<td>0.78</td>
<td>0.42</td>
<td>0.33</td>
<td>1507</td>
</tr>
<tr>
<td>Weighted average</td>
<td>0.77</td>
<td>0.44</td>
<td>0.34</td>
<td>1507</td>
</tr>
</tbody>
</table>

### Table IV: Classification Report for Set 2

<table>
<thead>
<tr>
<th>Class Label</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rub hands Palm to</td>
<td>0.89</td>
<td>0.88</td>
<td>0.88</td>
<td>460</td>
</tr>
<tr>
<td>Palm</td>
<td>0.91</td>
<td>0.39</td>
<td>0.54</td>
<td>510</td>
</tr>
<tr>
<td>Fingers Interlocked</td>
<td>0.57</td>
<td>0.90</td>
<td>0.70</td>
<td>505</td>
</tr>
<tr>
<td>Thumb Rub</td>
<td>0.72</td>
<td>0.72</td>
<td>0.72</td>
<td>1475</td>
</tr>
<tr>
<td>Micro average</td>
<td>0.79</td>
<td>0.72</td>
<td>0.71</td>
<td>1475</td>
</tr>
<tr>
<td>Macro average</td>
<td>0.79</td>
<td>0.72</td>
<td>0.70</td>
<td>1475</td>
</tr>
<tr>
<td>Weighted average</td>
<td>0.79</td>
<td>0.72</td>
<td>0.70</td>
<td>1475</td>
</tr>
</tbody>
</table>

Table III, IV is the classification report for Set 1 and 2 after the completion of the training process and the model is saved. Scikit-learn, an open source library for machine learning is used to compute the metrics such as precision, recall, f1 score and support. The precision is the ratio of TP / (TP + FP) where TP is the number of true positives and FP is the number of false positives. Precision is the ability of the classifier not to label a negative sample as positive. The recall is the ratio TP / (TP + FN) where TP is the number of true positives and FN is the number of false negatives. The recall is intuitively the ability of the classifier to find all the positive samples.

The F-beta score can be interpreted as a weighted harmonic mean of the precision and recall, where an F-beta score reaches the best value at 1 and worst score at 0.

### VII. Conclusion

Gesture recognition and interaction has remained a popular area of research with many creative discoveries established in the past. Motion based game controllers such as Leap Motion Controller, Microsoft Kinect can track hand and arm movements thereby providing gesture interaction in real-time. However, besides the entertainment industry and other industries, such as manufacturing and automation, gesture interaction is expanding in the healthcare sector for patients with mobility and other health-related problems. One example where the tracking and identification of gestures are of interest is the process of hand washing. The process of hand washing involves dynamic hand hygiene gestures. The paper presents ‘WHO-hand hygiene gesture classification system’ with preliminary results based on ResNet-50 framework, applied on hand hygiene video recordings. It is noted from the results that the system performs better in class prediction with higher number of training steps. However, increase in the training steps result in a longer period of training. The training time for the data in Set 2 with 50 epochs, has consumed 52 hrs. The accuracy level achieved was >70 %. The accuracy-speed trade-off is the key factor in deploying deep learning solutions. In future, more sophisticated workstation; NVIDIA GPU with
increased storage and memory will be utilized. Experiments will be conducted with larger data sets; recordings of the professional health care workers will be incorporated. ResNet-101, ResNet-152 with deeper layers will be explored to improve the performance of the system and to predict the hand hygiene stages in real-time. 

**Supplementary Materials:** The complete robust hand hygiene dataset recorded for 30 participants along with the python code; model and results for the following are available online:


**CONFLICT OF INTEREST**

The authors declare no conflict of interest.

**REFERENCES**


Rashmi Bakshi received the B.S in information technology from Eastern Mediterranean University, Northern Cyprus in 2011. She has received the masters in computer science from University of Bern, Switzerland in 2013. She has worked as an assistant lecturer at Vivekananda Institute of Professional Studies, India and taught computing related modules. (2013-2016). She has received her M. Phil in engineering from Technological University, Ireland in 2021. She is currently working as a part time assistant lecturer at the School of Electrical and Electronic Engineering, TU Dublin.