Viewpoint Invariant Convolutional Networks for Identifying Risky Hand Hygiene Scenarios

Michelle Guo[1] **Albert Haque**[1] **Serena Yeung**[1] **Jeffrey Jopling**[1]
Lance Downing[1] **Alexandre Alahi**[1,2] **Brandi Campbell**[3] **Kayla Deru**[3]
William Beninati[3] **Arnold Milstein**[1] **Li Fei-Fei**[1]

[1]Stanford University [2]École Polytechnique Fédérale de Lausanne [3]Intermountain Healthcare

Abstract

Hand hygiene is essential in preventing costly and potentially lethal hospital-acquired infections. Proposed by the World Health Organization, the Five Moments of Hand Hygiene is an evidence-based and field-tested categorization designed to measure and improve clinical hand hygiene. In this paper, we show the effectiveness of a viewpoint invariant convolutional network for automatically identifying the five moments of hand hygiene. Using de-identified images from privacy-safe depth sensors inside patient rooms, we demonstrate the viability of our deep learning method and provide interpretable visualizations. Our results show that computer vision can offer high-fidelity monitoring of this paramount clinician behavior.

1 Introduction

Hand hygiene is often the first line of defense in preventing hospital associated infections [1, 2]. A meta-analysis [3] estimated that hospital acquired infections cost the United States close to $10 billion each year. While it is one of the most important steps in the clinical care process, hand hygiene is notoriously difficult to track [4]. The auditing and observing of hand hygiene compliance, whether in an outpatient clinic or intensive care unit, requires extra staff and dedicated time.

To help combat hospital acquired infections, the World Health Organization proposed the Five Moments of Hand Hygiene [4] as a method of categorizing critical hand washing events. Because these events pose health risks to the patient, clinicians should wash their hands: (i) before touching a patient, (ii) before a cleaning or aseptic procedure, (iii) after exposure to a patient's bodily fluids, (iv) after touching a patient, and (v) after touching a patient's surroundings or environment [4].

Given the recent success of convolutional networks [5–9], they are uniquely poised to perform a more comprehensive appraisal of hand hygiene in the hospital setting. In previous work where deep learning was used in the rating of diabetic retinopathy [10] and dermatologic photos [11], static images were used as the training and evaluation sets. The act of hand hygiene, is dynamic: visitors and clinicians enter and interact with patient environments in many ways.

Existing technologies such as radio-frequency identification (RFID) systems are used to track people in hospitals [12] and are generally cheap and easy to deploy [13]. Despite these benefits, this generally requires that hospital staff wear special items to register their position during specific events (e.g., before entering a room [14, 15]). A more scalable solution to track humans with higher precision, at finer resolution and in a non-intrusive fashion is needed. Computer vision has been shown to provide fine-grained activity understanding in hospitals [16]. However, computer vision studies for hand hygiene are limited. In [17], the authors trained a classifier to detect antibacterial dispenser usage outside patient rooms. While this is a promising first step, their method does not detect physical contact with the patient and if used as an alarm system, is prone to many false positives.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

In this work, we demonstrate the effectiveness of a viewpoint invariant convolutional network at identifying instances of the five moments of hand hygiene. We show how a modern deep network architecture can be applied to depth-based imaging sensors (e.g., Kinect) to detect when a clinician physically touches the patient or environment. We evaluate our method on real data collected from an intensive care unit. Additionally, we provide interpretable 3D visualizations to understand our model's spatial reasoning.

2 Method

Data. Seven depth sensors were installed on the ceiling inside patient rooms in an intensive care unit. Each depth sensor reports 3D point cloud data in the form (x, y, z) where x, y, z are values in real-world meters with respect to the camera's coordinate system. Cameras were manually calibrated and all points were projected onto a shared coordinate system. A total of 20,038 annotated examples were collected over a three month period. Each annotation contains a multi-class label: touching the patient, touching the environment, or no action. Table 1 shows dataset statistics.

	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9
# Environment Touch	85	621	928	536	294	0	9	604	4
# Patient Touch	542	2,265	4,626	1,723	1,470	891	0	1,691	0
# Frames	872	7,391	12,196	6,519	3,605	1,059	82	3,225	59
Sensors (Viewpoints)	A	A-B	A-F	B	C	D	E	F	G

Table 1: Specifications of the data subsets used in our experiments. For each dataset we report the number of annotated environment touch frames, patient touch frames, and total number of frames. We assign the letters A-G for our 7 sensors. All sets have one sensor each except for Sets 2 and 3 which contain 2 and 6 sensors, respectively. The motivation of multiple subsets is to evaluate our model's ability at learning single and multiple viewpoints.

Viewpoint Invariance. Our goal is to train a model capable of classifying the five moments of hand hygiene in hospitals worldwide. To achieve this goal, our model must generalize across hospitals and the differing viewpoints it may encounter in different patient rooms. Following the work of [17], we use a spatial transformer network [18] to help our model generalize across viewpoints.

The first layer of our convolutional network is a spatial transformer. The purpose of the spatial transformer is to project the input point cloud onto a viewpoint invariant feature space. Our model internally predicts two-dimensional[1] transformation parameters θ, which are needed to transform the input. This produces the affine transformation matrix:

$$A_\theta = \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \\ \theta_4 & \theta_5 & \theta_6 \end{bmatrix} = \begin{bmatrix} s_x & \varphi_x & t_x \\ \varphi_y & s_y & t_y \end{bmatrix}. \tag{1}$$

This particular transformation allows for cropping or translation via t_x, t_y, scaling via s_x, s_y and rotation or shearing via φ_x, φ_y. We use a bilinear sampling kernel, which produces the output V:

$$V_i^c = \sum_n^H \sum_m^W X_{nm}^c \max(0, 1 - |x_i^s - m|) \max(0, 1 - |y_i^s - n|) \tag{2}$$

where X is the input point cloud rendered as a 2D image, (x_i^s, y_i^s) are source coordinates in the input X that define the sample points, and where the superscript c denotes the channel index. The output (and now invariant) feature map V is then fed to a densely connected convolutional network.

Classification Network. Given the recent success of skip connections [8] in convolutional neural networks, we use a densely connected convolutional neural architecture [7] (DenseNet) to perform classification of the five moments activities. The input to the DenseNet is the viewpoint invariant feature map V from the spatial transformer network.

[1] Although learned 3D transformation parameters have been proposed [18–20], we opt for a two-dimensional transformation due to its fast runtime performance, mathematical and implementation simplicity.

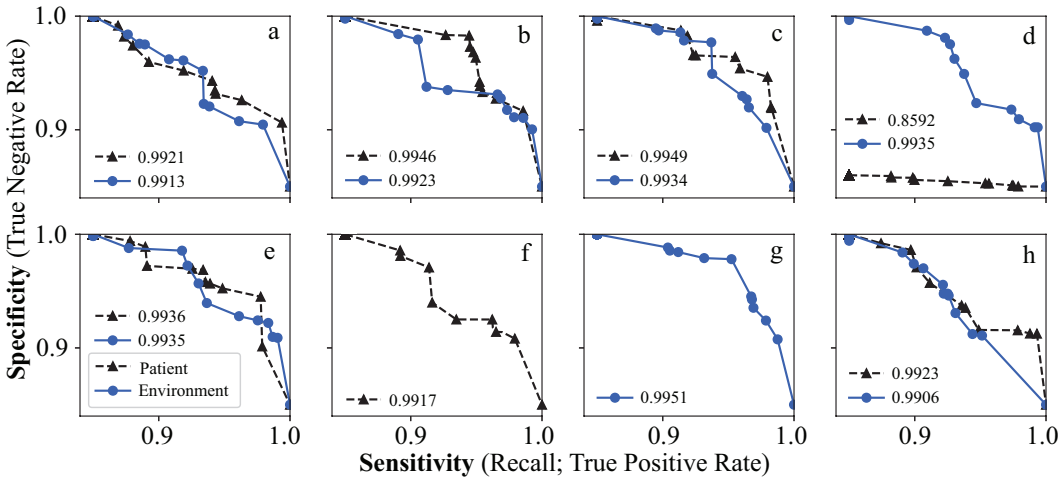

Figure 1: Sensitivity versus specificity for different experiments. Each subfigure (a)-(h) describes a different experiment, run on validation sets for sets 1 through 8, respectively. Dashed black lines indicate the patient-touch classifier and blue solid lines indicate the environment-touch classifier. Numeric values in the bottom left of each plot denote area under the curve. Subfigures (f) and (g) only have one experiment because the dataset contains zero examples a contact task (see Table 1).

Physical Contact With	STN	Set 1	Set 2	Set 3	Set 4	Set 5	Set 6	Set 7	Set 8	Set 9
Patient	No	93.1	**95.1**	**94.1**	93.9	95.8	89.6	—	95.4	—
	Yes	**97.7**	93.8	92.4	**95.8**	**98.6**	**93.9**	—	**97.7**	—
Environmental	No	93.7	**93.7**	93.4	92.5	94.6	—	93.8	90.4	91.7
	Yes	**95.4**	91.8	**94.1**	**93.0**	**94.7**	—	93.8	**92.2**	91.7

Table 2: Accuracy with and without the spatial transformer network. The patient and environmental touching classifiers were evaluated on individual and combined sensor views (see Table 1). STN denotes the spatial transformer network. A dash denotes datasets with zero positive examples.

Optimization Objective. Our dataset contains a large number of negative examples (70%) where there is no patient contact nor environmental contact. Class-balancing addresses this problem: the ratio of negative to positive examples R is fixed such that training and testing R are equivalent. However, this approach discards many negative examples that could be useful signals for learning a more robust model (e.g., hard negatives). An alternative approach which avoids discarding examples is to weight positives and negatives such that the loss weights are inversely proportional to their frequency in the training set [21]. Instead, we use:

$$\mathcal{L}(p_t) = -(1-p_t)^{\gamma} \log p_t \quad \text{where} \quad p_t = \begin{cases} p, & \text{if } y = 1 \\ 1-p, & \text{otherwise} \end{cases} \tag{3}$$

where p is the post-softmax value for any one class prediction and γ scales the weighting of very confident predictions. This loss function is known as the Focal Loss [22]. It dynamically adjusts the loss weighting to give more weight to hard negatives during the entire training period.

3 Experiments & Discussion

Figure 1 shows the sensitivity versus specificity of our method for classifying patient contact and environmental contact. When training one model per viewpoint, our model classifies patient and environment touching well. Patient contact classification performs slightly better than environmental contact classification. This is potentially due to the fact that clinicians are generally in fixed positions when touching the patient (e.g. hovering over the bed), while the locations and body poses of clinicians during environmental contact can be highly variable. When trained and evaluated on all sensor viewpoints (set 3), our model is nearly equally performant.

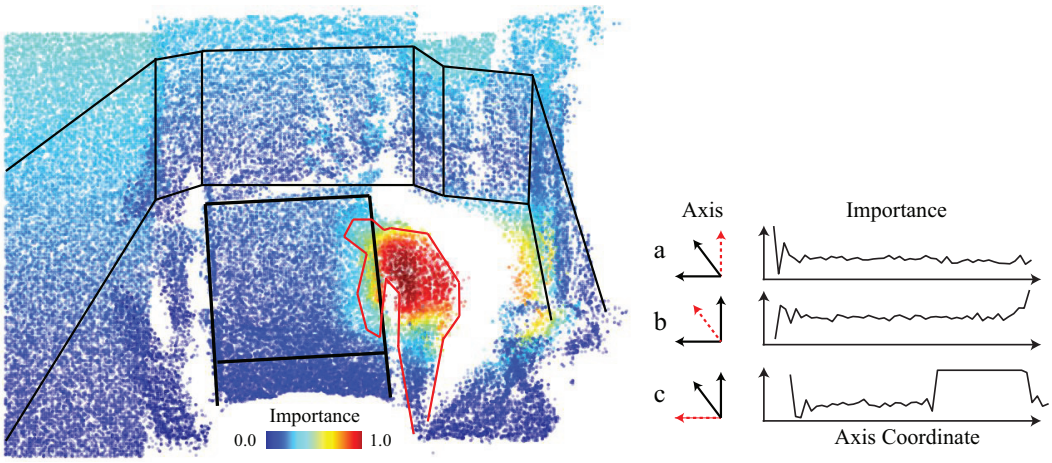

Figure 2: (left) Attention map over the physical space. The point cloud depicts a scene captured from our depth sensor. To assist the reader, black lines were drawn to denote the geometry of the patient's room (including the bed) and the red lines outline the clinician. Additionally, points higher in the physical space (i.e., toward the ceiling) have a lighter blue-green color; this is not necessarily indicative of importance. Red points indicate spatial regions containing high signal for discerning physical contact with a patient. The model learns to focus on the the contact point between the person and the patient bed when classifying patient touch. **(right) Importance magnitudes collapsed onto a single spatial dimension.** Subfigures (a)-(c) show histograms of importance values for a single axis. The axis is denoted by the red dashed arrow in the three-axis diagram.

Viewpoint Invariance. We compare the performance of training one model per viewpoint compared to training a single model for all viewpoints. To measure this, we compute a weighted accuracy over all viewpoint models using dataset sizes. Training separate patient contact classifiers for each sensor achieves a weighted accuracy of 96.9%, whereas the single model variant achieves 94.1%. For classifying environmental contact, separate models per viewpoint achieves 93.1% while our single model achieves 94.1%. This demonstrates the viewpoint invariance of our model: training a single model on multiple views is competitive with per-sensor models. There are many ways for a clinician to come in contact with the environment. It is likely that our model is able to leverage the diverse environmental contact examples from other rooms and viewpoints to improve performance. However, contact with a patient is much less variable. A clinician generally hovers over the bed when touching the patient. This benefits models optimized for a single viewpoint. Table 2 shows the results of our model with and without a spatial transformer module on different subsets. The results show that our model performs better overall with the spatial transformer network.

Spatial Attention. To better understand the model's predictions, we performed an occlusion mask analysis [23] on a model trained on predicting physical contact with the patient. This is performed by occluding (i.e., zero-padding) the input with variable sized windows and monitoring the model's confidence (i.e., post-softmax result). This process is repeated across different regions of the point cloud tens of thousands of times. The result is shown in Figure 2. Red points indicate input regions important for correctly classifying the scene as contact being made with the patient. The model correctly localizes the clinician in the room and assigns a high importance to the clinician since he or she is leaning over the bed with his or her arm extended.

4 Conclusion

In this paper, we showed the effectiveness of a viewpoint invariant convolutional network at automatically identifying instances of the five moments of hand hygiene. We collected point cloud data from depth sensors installed in an intensive care unit. When evaluated with a spatial transformer module, our model is able to perform well given multiple sensor viewpoints across the hospital unit. We presented intuitive and interpretable 3D visualizations of our model's predictions which can be used to educate and improve clinician behavior. This work shows that computer vision can offer automatic and high-fidelity monitoring of the five moments of hand hygiene.

References

1. Pittet, D. *et al.* Effectiveness of a Hospital-Wide Programme to Improve Compliance with Hand Hygiene. *The Lancet* **356,** 1307–1312 (2000).
2. Magill, S. S. *et al.* Multistate Point-Prevalence Survey of Health Care–Associated Infections. *New England Journal of Medicine* **370,** 1198–1208 (2014).
3. Zimlichman, E. *et al.* Health Care–Associated Infections: A Meta-Analysis of Costs and Financial Impact on the US Health Care System. *JAMA Internal Medicine* **173,** 2039–2046 (2013).
4. Organization, W. H. in, 100–101 (2009).
5. LeCun, Y. & Bengio, Y. Convolutional Networks for Images, Speech, and Time Series. *The Handbook of Brain Theory and Neural Networks* **3361,** 1995 (1995).
6. Krizhevsky, A., Sutskever, I. & Hinton, G. E. *Imagenet Classification with Deep Convolutional Neural Networks* in *NIPS* (2012), 1097–1105.
7. Huang, G., Liu, Z., Weinberger, K. Q. & van der Maaten, L. Densely Connected Convolutional Networks. *CVPR* (2017).
8. He, K., Zhang, X., Ren, S. & Sun, J. *Deep Residual Learning for Image Recognition* in *CVPR* (2016), 770–778.
9. Fout, A., Shariat, B., Byrd, J. & Ben-Hur, A. *Protein Interface Prediction using Graph Convolutional Networks* in *NIPS* (2017).
10. Gulshan, V. *et al.* Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316,** 2402–2410 (2016).
11. Esteva, A. *et al.* Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. *Nature* **542,** 115–118 (2017).
12. Fuhrer, P. & Guinard, D. *Building a Smart Hospital Using RFID Technologies: Use Cases and Implementation* (Department of Informatics-University of Fribourg, 2006).
13. Coustasse, A., Tomblin, S. & Slack, C. Impact of Radio-Frequency Identification (RFID) Technologies on the Hospital Supply Chain: A Literature Review. *Perspectives in Health Information Management* **10** (2013).
14. Simmonds, B. & Granado-Villar, D. Utility of an Electronic Monitoring and Reminder System for Enhancing Hand Hygiene Practices in a Pediatric Oncology Unit. *American Journal of Infection Control* **39,** E96–E97 (2011).
15. Yao, W., Chu, C.-H. & Li, Z. *The Use of RFID in Healthcare: Benefits and Barriers* in *International Conference on RFID-Technology and Applications* (2010), 128–134.
16. Ma, A. J. *et al.* Measuring Patient Mobility in the ICU Using a Novel Noninvasive Sensor. *Critical care medicine* **45,** 630–636 (2017).
17. Haque, A. *et al.* Towards Vision-Based Smart Hospitals: A System for Tracking and Monitoring Hand Hygiene Compliance. *Machine Learning in Healthcare Conference* **68** (2017).
18. Jaderberg, M., Simonyan, K., Zisserman, A. & Kavukcuoglu, K. *Spatial Transformer Networks* in *NIPS* (2015), 2017–2025.
19. Bas, A., Huber, P., Smith, W. A., Awais, M. & Kittler, J. 3D Morphable Models as Spatial Transformer Networks. *arXiv:1708.07199* (2017).
20. Bhagavatula, C., Zhu, C., Luu, K. & Savvides, M. Faster Than Real-Time Facial Alignment: A 3D Spatial Transformer Network Approach in Unconstrained Poses. *arXiv:1707.05653* (2017).
21. He, H. & Garcia, E. A. Learning from imbalanced data. *Transactions on Knowledge and Data Engineering* **21,** 1263–1284 (2009).
22. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal Loss for Dense Object Detection. *arXiv:1708.02002* (2017).
23. Zeiler, M. D. & Fergus, R. Visualizing and Understanding Convolutional Networks. *ECCV,* 818–833 (2014).