Vision-Based Hand Hygiene Monitoring in Hospitals

CS229 Fall 2015 Project Report
Zelun Luo, Boya Peng, Zuozhen Liu

Abstract

Hand hygiene has been shown to be an effective intervention to reduce transmission and infections in many studies. This project focuses on interpreting visual clinical data for hand hygiene monitoring. We propose two distinct deep learning approaches to detect hand hygiene action on manually collected and labeled data. Specifically, we investigate a fixed-window and a pose-based hand detector using Convolutional Neural Network (CNN). We show both approaches are able to achieve high accuracy and outperform our baseline model using linear Support Vector Machine (SVM) classifier.

1 Introduction

With recent success of deep learning in computer vision, visual clinical data can be exploited to improve the understanding of patient experience and environment during hospital stays. Such data can contain rich information about patient condition such as the appearance of distress, which has been described as the 6th vital [1], as well as details about the occurrence and nature of clinical activities ranging from patient care to bundle compliance and hand hygiene. However, visual clinical data still remains an underexplored source of information in the healthcare settings.

In this project, we want to make use of valuable visual clinical data in the hand hygiene setting where the objective is to detect when a person performs the hand hygiene action using various computer vision and machine learning methods. This action is defined by a person placing his or her hand under a hand hygiene dispenser and receiving soap. Therefore, this problem can be formulated as a supervised binary classification task with raw visual image inputs.

Provided that deep learning methods have achieved state-of-the-art performance in various computer vision tasks in recent years, we want to apply CNN to solving this classification task. We are also interested in exploring pose-based approaches that may be easily extended to monitoring clinical activities in the healthcare settings.

2 Dataset

For an initial pilot study, we collect both depth and RGB data from a depth sensor mounted in a lab environment. We collect a dataset of 2-hour depth and RGB signals, from which we extract 25,465 frames containing 630 positive hand hygiene frames. Examples of depth images are shown in Fig 1. These frames are then used to train and evaluate our SVM baseline model, CNN-based hand hygiene detection model, and the pose-based model. Since the dataset is highly imbalanced as we have far more negative frames than positive ones, we use cross validation to tune the ratio of positive and negative frames in the training set.

For our pose-based approach, we use the hand dataset found at Oxford Visual Geometric Group's repository [2]. We use 11,700 hand images with 12,800 synthesized negative examples to train the hand classifier.

Figure 1: Examples of depth images from our dataset. From left to right, the first two are positive instances of hand hygiene, and the last two are challenging negative instances.

3 Approach

3.1 SVM baseline model

For each image, we first crop a 64x64 region that contains the dispenser. This region encodes rich information regarding the hand hygiene action and we can train a classifier using features directly based on raw pixel values. Our classifier is a linear SVM model with Hinge loss function. In order to combat overfitting, we also incorporate L1 regularization into our SVM model which is equivalent to the optimization problem expressed below.

$$\begin{aligned} \min_{\gamma,\mathrm{w},\mathrm{b}} \quad & \frac{1}{2}\|w^2\| + C\sum_{i=1}^{m}\xi_i \\ s.t. \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \\ & i = 1, \ldots, m \end{aligned}$$

3.2 Fixed window hand hygiene detection

In this approach, we train a CNN model to detect whether hand hygiene action occurs in a video frame. We select the fixed window to be a cropped region containing the dispenser (a 64×64 pixles region near the dispenser).

The network architecture consists of two convolutional layers, each followed by a max pooling layer, and two fully connected layers. Since the input images have relative small dimension, we decide that two convolution steps are enough to extract important high-level feature representations for classification. The output from the fully connected layer is a binary classification of whether the hand hygiene action is occurring, and we optimize a logistic loss function using stochastic gradient descent. We then use cross validation to tune all hyperparameters to determine the final architecture of our CNN-based model shown in Fig 2.

In our experiments, we compare this model against a similar CNN model trained on the full frame image (320×240 pixels). In addition, we also compare the accuracy between models trained on RGB images and depth images respectively.

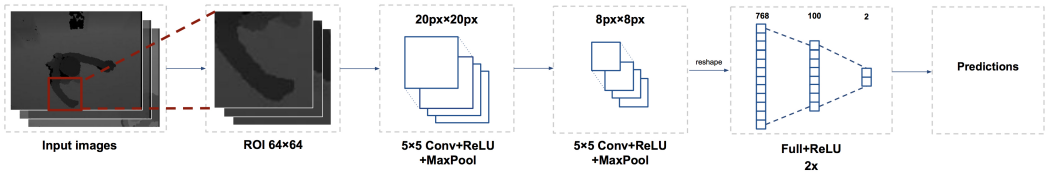

Figure 2: CNN-based hand hygiene detection model architecture

3.3 Pose-based approach

We train a CNN-based detector for hand joint in each detected human region, since the hand is what performs the hand hygiene action. We then consider hand hygiene to be performed if a hand is detected in the physical space immediately under the hand hygiene dispenser. In such scenario, the hand placement is most likely to trigger the soap to be dispensed. In this approach, we first train a CNN-based hand detector that detects hands in a 32×32 pixels region. The model consists of two convolutional layers, each followed by a max pooling layer and the training procedure is almost identical to the first CNN model.

In the classification phase, for each input image, we use the sliding window method to extract regions of size 32×32 pixels with a stride of 4 pix-

els. We then feed each extracted region into our hand detector and predict if a hand is detected. Hand hygiene is considered to be performed if the distance between the closest hand and the dispenser is smaller than a fixed threshold. Fig 3 shows the architecture of our pose-based model.

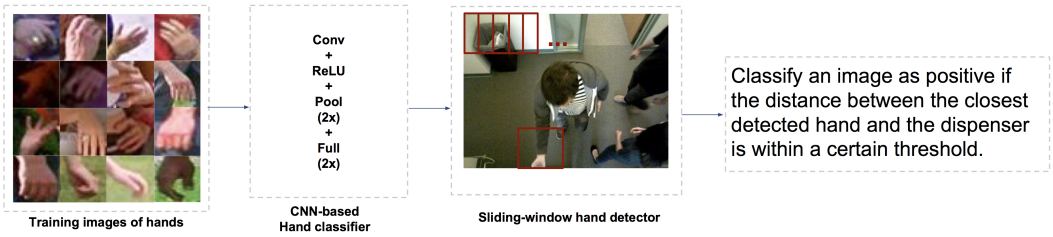

Figure 3: pose-based model architecture

4 Experiments

Using the dataset described in section 2, we perform evaluations on all of the approaches discussed in previous section on both RGB and depth images. Taking the target class imbalance into account, we use Mean Average Precision(MAP) as our metric to reflect the performance of our different detectors. The results for these detectors are displayed in Table 1. For privacy reasons, depth images are preferred in realistic settings and we want to make sure that our models are able to adapt to this challenge. Note that our hand detector in pose-based approach is trained only on RGB images; therefore, this approach is currently only limited to RGB images.

From Table 1, we observe that both approaches outperform linear SVM baseline on RGB images. The intuition is that linear SVM baseline imposes strong bias on the hypothesis class in RGB space and the model is underfitting due to lack of complexity. One promising solution is to train SVM with various kernels to learn nonlinear boundaries in the RGB space. However, the SVM baseline performs surprisingly well on depth images. We believe that the reduction in dimension from depth images gives rise to a simpler decision boundary that can be well represented by a linear boundary.

The CNN model over dispenser region is able to achieve strong performance on both RGB and depth images. The results also validate the success of CNNs achieving state-of-the-art performance in vision tasks. However, the results obtained from CNN over the entire image are less satisfactory due to undesired noise in the rest part of the image. Fig. 4 shows the detection results on cropped depth images. We observe that the CNN model over dispenser region correctly detects hand hygiene action when an arm is stretching out to the dispenser. However, from the false negative results shown in the second row, this model fails to detect the action when a person comes too close to the dispenser. This limitation is caused by partial occlusions due to a top-down viewing angle of depth sensor.

The pose-based approach has a worse performance than fixed-window CNN. One challenge is that people's hands are often occluded by the dispenser when they are performing hand hygiene and this method depends on hand detection to make final prediction. Such input frames are likely to be classified as negative as the hand detector is not able to detect any hands. On the other hand, false positive results sometimes occur when people's hands are within the distance threshold from the dispenser but are not performing hand hygiene actions. However, this method has a potential benefit of being able to tie the action to its performer. Overall speaking, this model achieves decent performance considering its simplicity. More complex pose-based models can be further explored to resolve the challenges mentioned above. Fig. 5 shows different detection results using the pose-base approach.

	RGB	Depth
SVM baseline over dispenser region	0.561	0.889
CNN over full image	0.695	0.450
CNN over dispenser region	**0.956**	**0.937**
Pose-base approach	0.807	–

Table 1: Average precision of hand hygiene action detection.

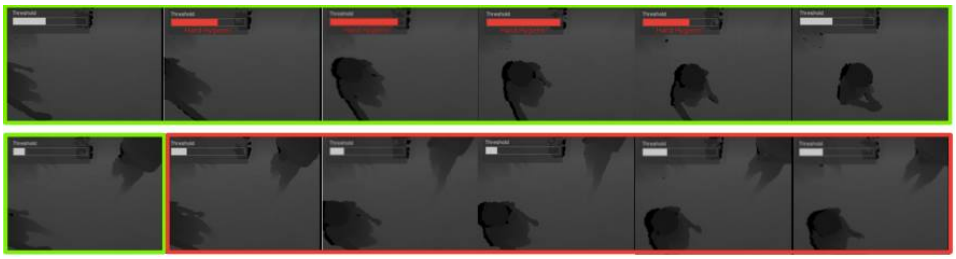

Figure 4: Examples of detection results using CNN approach over dispenser region. Green for correct labelings; red for incorrect labelings.

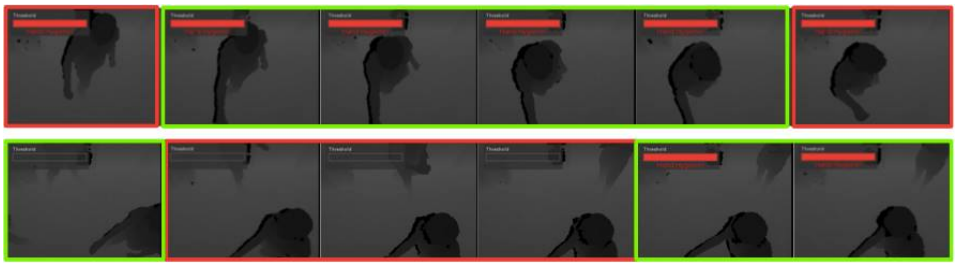

Figure 5: Examples of detection results using pose-based approach. Green for correct labelings; red for incorrect labelings.

5 Conclusion

We believe that the recent success of using machine learning techniques over depth signals to perceive the world could have an unprecedented impact in health care. We have shown that it is possible to detect hand hygiene compliance, an important component of reducing the cost associated with hospital-acquired infections.

For future work, we plan to work on a pose-estimation model which is invariant to different view points and can handle self-occlusion. A pose-estimation model allows more general applications such as person identification, activity recognition and characterization, which can be broadly applied in a healthcare setting. Although there has been various work focusing on 2D/3D pose estimation from RGB images [4] [5], work on 3D pose estimation from depth images are relatively scarce. Shotton *et al* [3] appear to represent the state of the art for pose estimation from depth images, but they still fail to address the challenges of handling different camera view points (e.g. top view) and occlusions.

References

[1] D. Howell and K. Olsen. Distress the 6th vital sign. *Current oncology*, 18(5):208, 2011.

[2] A. Mittal, A. Zisserman, and P. Torr. Hand detection using multiple proposals.

[3] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. *Real-time human pose recognition in parts from a single depth image*. In Proc. CVPR. IEEE, 2011.

[4] J. Tompson, A. Jain, Y. LeCun, and C. Bregler. *Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation*.

[5] C. Wang, Y. Wang, Z. Lin, A. L, Yuille, and W. Gao. *Robust estimation of 3d human poses from a single image*. in IEEE Conference on Computer Vision and Pattern Recognition, 2014.