Hand hygiene monitoring based on segmentation of interacting hands with convolutional networks

Armin Dietz, Andreas Pösch, Eduard Reithmeier

Armin Dietz, Andreas Pösch, Eduard Reithmeier, "Hand hygiene monitoring based on segmentation of interacting hands with convolutional networks," Proc. SPIE 10579, Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications, 1057914 (6 March 2018); doi: 10.1117/12.2294047

SPIE.

Event: SPIE Medical Imaging, 2018, Houston, Texas, United States

Hand hygiene monitoring based on segmentation of interacting hands with convolutional networks

Armin Dietz[a], Andreas Pösch[a], and Eduard Reithmeier[a]

[a]Leibniz Universität Hannover, Nienburger Str. 17, Hannover, Germany

ABSTRACT

The number of health-care associated infections is increasing worldwide. Hand hygiene has been identified as one of the most crucial measures to prevent bacteria from spreading. However, compliance with recommended procedures for hand hygiene is generally poor, even in modern, industrialized regions. We present an optical assistance system for monitoring the hygienic hand disinfection procedure which is based on machine learning. Firstly, each hand and underarm of a person is detected in a down-sampled 96 px x 96 px depth video stream by pixelwise classification using a fully convolutional network. To gather the required amount of training data, we present a novel approach in automatically labeling recorded data using colored gloves and a color video stream that is registered to the depth stream. The colored gloves are used to segment the depth data in the training phase. During inference, the colored gloves are not required. The system detects and separates detailed hand parts of interacting, self-occluded hands within the observation zone of the sensor. Based on the location of the segmented hands, a full resolution region of interest (ROI) is cropped. A second deep neural network classifies the ROI into ten separate process steps (gestures), with nine of them based on the recommended hand disinfection procedure of the World Health Organization, and an additional error class. The combined system is cross-validated with 21 subjects and predicts with an accuracy of 93.37% (± 2.67%) which gesture is currently executed. The feedback is provided with 30 frames per second.

Keywords: Hand hygiene, segmentation, machine learning, hand tracking, gesture recognition

1. INTRODUCTION

The World Health Organization (WHO) suggests a hand disinfection process consisting of six steps, which are represented in Figure 1.[1] Each step consists of unique gestures which have to be applied in a rubbing movement. As some of the steps have to be performed in variation to cover the areas of both hands, the whole procedure can be classified into nine separate gestures. The recommended duration of the entire procedure is 20-30 seconds.

Figure 1. Six steps of the hand disinfection process. Some steps must be performed in variation to cover the areas of both hands.

Further author information:
Armin Dietz: E-mail: armin.dietz@imr.uni-hannover.de, Telephone: +49 511 762 5817

Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications, edited by Jianguo Zhang, Po-Hao Chen, Proc. of SPIE Vol. 10579, 1057914 · © 2018 SPIE
CCC code: 1605-7422/18/$18 · doi: 10.1117/12.2294047

To monitor the compliance, the most common methods are direct observation of practice, self-report of health-care workers and indirect calculations, which can be based on the usage of hand disinfection solution or by measurements of electronic devices.[2] Out of those methods, direct observation is considered as the 'gold standard' and it is the only method, that not only evaluates the quantity of the hand disinfection process, but also the quality. Nonetheless, direct observation also has its disadvantages, such as high labor intensity and labor cost, untrained personnel that is not always trained in the same manner on how to monitor the hand disinfection process, and the so-called Hawthorne-Effect, which states that people behave differently when they know that they are being observed.[2] To reduce the impact of the Hawthorne-Effect, video observation systems have been tried that evaluate the quality of the process automatically by detecting if required gestures have been performed. One commercially available system is SureWash, which uses a RGB camera.[3] However, it requires a fixed background and has currently only been deployed for training purposes and not for real-time monitoring. Another system, that uses a depth sensor similar to the system presented in this work is WashInDepth.[4] It detects hands by filtering background pixels that have been static when no person has been in the frame. Although this system works well under certain conditions, this background removal will likely not work when there are several moving people inside the observation area or when the hands are close to the body. We propose a monitoring system that detects specifically hands and underarms by pixelwise classification. This allows for a more robust region of interest detection for further gesture classification. To acquire the required amount of labeled training data, we present a novel method based on color segmentation of colored gloves, that is quick to implement.

2. MATERIALS AND METHODS

We use a Microsoft Kinect v2 sensor with the open source driver libfreenect2 to acquire a 512 px x 424 px resolution depth and infrared stream, and a rgb color stream that is registered to them.[5] The registration is based on calibration parameters that are provided by the libfreenect2 software. The sensor is facing downwards in a slight angle from two meters height, observing the scene around a disinfection dispenser, as it is illustrated in Figure 2. The monitoring of the hand disinfection process is based on two steps, a pixelwise segmentation of the separate hands and underarms, and a gesture recognition. We use the Keras deep learning library with the Tensorflow backend for training of neural networks.[6]

Figure 2. Illustration of the setup in practice, with a time-of-flight sensor facing downwards observing the scene around a disinfection dispenser.

2.1 Pixelwise segmentation of interacting hands

To achieve a detailed segmentation of each hand and underarm, we use a fully convolutional neural network for pixelwise classification. As labeling detailed hand parts for training data manually is very time consuming, we propose a method to automatically create labels.

2.1.1 Data Acquisition

The subjects wear colored gloves with separate colors for each hand and underarm. We record 67375 frames of the depth, the infrared and the registered color stream of eight subjects, that perform random movements with their hands, including hand disinfection process steps. The different colors of the gloves are then segmented using experimentally determined thresholds in the HSV color space. It is important to ensure that the subjects do not wear any similar colored clothes to the gloves. Furthermore any pixels further away than 160 cm from the sensor are set to black and thus any distracting background color is removed. Occurring noise of the depth signal, that is especially apparent on edges of objects due to the time-of-flight technology of the sensor, is reduced by removing all frames that are below a intensity of 2000 in the infrared image. The infrared image displays the intensity of the active infrared light in a range between 0 and 65536. The hands, with or without gloves, are always above the chosen threshold of 2000. A diffused LED-array on top and white reflectors on the sides help to achieve constant lighting conditions in the region of interest and thus make it possible to use constant threshold values. Without the additional reflectors on the sides, the automatic exposure correction of the Kinect sensor is overexposing the area of the hands. This is the case as most of the background is less illuminated than the area of the hands that is closer to the sensor. As the color stream is registered to the depth stream, we can label each pixel of the depth stream as either left hand, right hand, left underarm, right underarm, or background, depending on its color value. Figure 3 illustrates the color gloves and the automatically labeled hand mask.

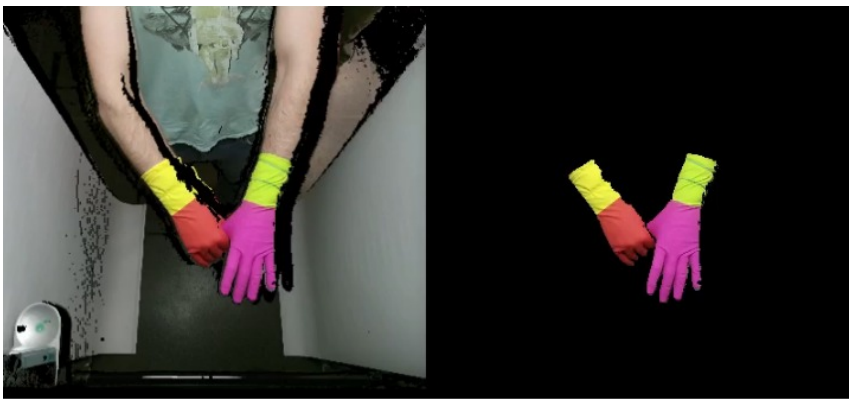

Figure 3. A subject is wearing colored gloves (left). The color frame is registered on the depth frame. Registration errors can be noted. The white reflectors on the sides help to achieve constant lighting conditions. Color thresholded hands, where each hand and underarm receive a label (right).

2.1.2 Network Architecture

The network for pixelwise classification is a variation of the U-net, proposed by Ronneberger et al.[7] It has a encoding and a decoding part. The encoding part consists of four blocks. One block contains two 3x3 convolutions (zero padded, except for the first block) and one 3x3 convolution with stride 2 for downsampling. Each layer is followed by an ELU activation.[8] At the end of each block is a dropout of 0.5. The number of filters is doubled every block, starting with 32. Every block in the decoder part consists of a 2x2 deconvolution layer with stride 2 that upsamples the feature maps. These feature maps are concatenated with feature maps of the encoding part that have the same dimensions. Two 3x3 convolutions with ELU activations and a dropout of 0.5 follow. At the end, a 1x1 convolutional layer with 5 filters for the 5 classes is activated with a softmax function for classification.

2.1.3 Training

The neural network is trained for 50 iterations of all frames with Adam optimization.[9] We use a batch size of 32 and categorical cross entropy as a loss function. The frames are preprocessed differently before every new iteration for a higher variety of training data to prevent overfitting. The depth frame and the label mask are cropped randomly to 400 px x 400 px and then resized to 100 px x 100 px for a lower impact on hardware

requirements. Nearest neighbor interpolation is used to ensure that values of the label mask do not change by the interpolation. The frames are further randomly rotated by -25 to 25 degrees and shifted by -25 px to 25 px in both the horizontal and vertical direction. Each depth frame is normalized. We calculate the mean and standard deviation of each frame, excluding the pixels with value zero. Each frame is then subtracted by the mean and divided by the standard deviation. Excluding the zero pixels causes the area of the hands to be centered close zero, whereas it would be mostly the prevailing background area centered around zero when calculating the mean and standard deviation of the whole frame. Figure 4 shows a histogram of both cases. By calculating the mean and standard deviation for each frame individually, we achieve less variance of the input data against changing distances of the hands from the sensor. The output of the network is a pixelwise labeled image of 96 px x 96 px.

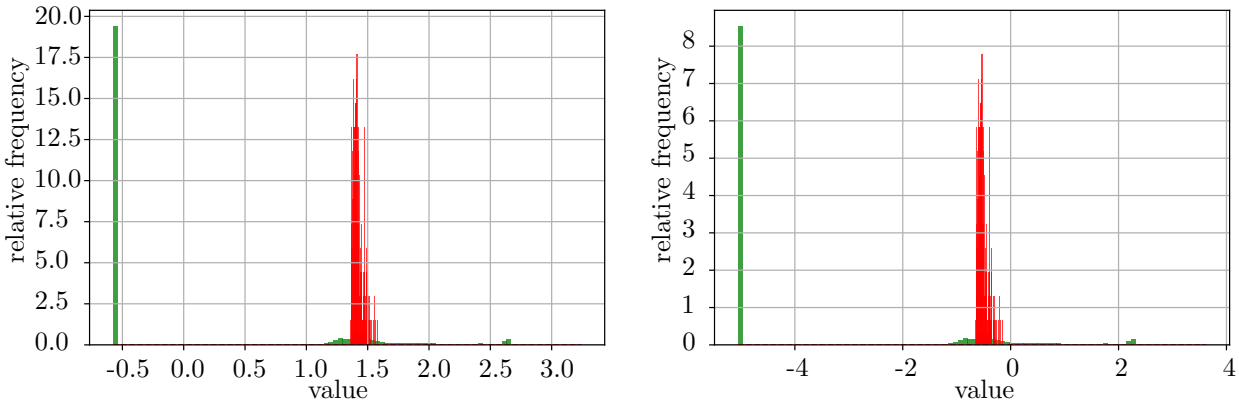

Figure 4. Histograms showing the distribution of values for normalized data where the mean and standard deviation are calculated from all data (left), and where they are calculated without zero pixels (right). All pixels that are not hands are green, the hand pixels are red. By excluding the zero pixels from the calculations, the values of the hand pixels are closer to 0, while the less relevant values of the background pixels (zero pixels) are further apart (note the different range of values in the histograms).

2.2 Gesture recognition

For gesture classification, we use a relatively small convolutional neural network with only 939912 parameters, which is a downsized version of the proposed network by Zeiler.[10] We record 83000 depth and infrared frames of 21 subjects, who perform the nine gestures of the hand disinfection process and an additional random gesture. All frames are labeled manually based on the performed gesture.

2.2.1 Region of Interest

The detailed pixel labels of the segmentation network allow for consistent inputs for the gesture classification network. Firstly, we calculate the bounding box of the segmented hands, not including the underarms. The bounding box is then upscaled to match the region of the hands in the depth and infrared frame of the original high resolution. Finally, all pixel that are not part of the hand or underarm labels inside the crop are set to zero and the frame is resized to 96 px x 96 px, which is close to its original shape.

2.2.2 Architecture of gesture recognition network

The gesture recognition network consists of 8 blocks. The input is convolved with 48 filters of size 7x7 and a stride of 2 in both directions. After an activation with a rectified linear unit (ReLU) follows a 2x2 max pooling with stride 2 and a batch normalization. The next block is similar, with 128 filters and a 5x5 convolution. The two following blocks consist each of 196 filters of 3x3 convolutions with stride 1 and a ReLU activation. The last convolution block has 48 3x3 filters with stride 1, a ReLU activation, and a 2x2 max pooling layer with stride 2. It follow 2 blocks of 0.5 dropout, a dense layer of 196 neurons and a ReLU activation. The output layer consists of 10 dense neurons for the 10 classes with softmax activation.

2.2.3 Training of gesture recognition network

The input of the network consists of two channels, the depth channel and the infrared channel. The input is normalized, with values higher than zero ranging between -1 and $+1$, and zero values set to -1.1. This is to separate the background from the hands and has achieved better results than normalizing all pixels to -1 and $+1$. The network is trained for 40 iterations of the training data set using stochastic gradient descent and a learning rate of 0.01, progressively divided by 0.5 every ten iterations. The gesture classification is validated using k-fold cross-validation, with 21 data sets and 7 folds. Thus, the network is trained with 18 data sets and validated with 3 different data sets. Each data set contains frames from different subjects.

3. RESULTS

Even though the input resolution for the segmentation network is very low with 100 px x 100 px, the network is able to detect detailed fingers and hand areas, as can be seen in Figure 5. The network also detects the fingers, when they are at the far edges of the frame and even when the hands are partly out of the frame. The network detects hands correctly when there are several persons in the frame. However, it is not possible to match the hands to the corresponding person. Left and right hand are differentiated well as long as both hands are inside the frame. The detailed information make it possible to create consistent crops that are necessary for the second neural network for a good gesture classification. The mean accuracy of the 7 validated folds of the gesture recognition network is 93.37% with a standard deviation of $\pm$ 2.67%. The segmentation network requires less than 10 ms for inference. We used a Intel Core i7-4770K CPU with a NVIDIA GeForce GTX960 GPU. The total prediction time is limited by the frame rate of the Kinect sensor with 30 frames per second.

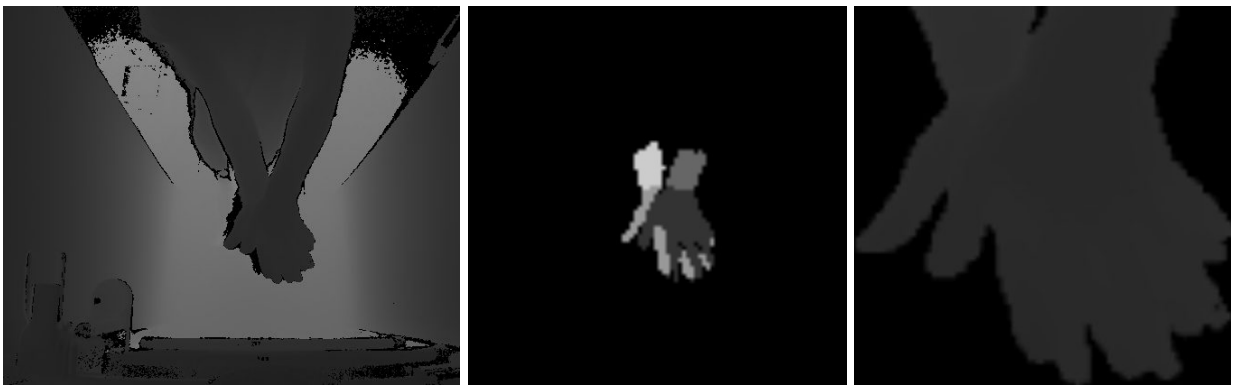

Figure 5. Input depth frame of a subject who is not wearing any gloves (left). Predicted labels for each hand and underarm in different gray scales (middle). Depth crop of the full resolution hands, that serves as input for the gesture recognition network (right).

4. CONCLUSIONS

We presented an approach to monitor if hand disinfection process steps are performed. By segmenting each hand and underarm separately, it is possible to receive robust data for a gesture classification network. We presented a novel technique to record labeled data automatically that is quick to implement. In future work, the output resolution of the segmentation network should be improved. This labeled output could then be used as an additional channel for the gesture recognition network which might improve the recognition accuracy. Both networks could also be combined into a single network.

REFERENCES

[1] Allegranzi, B., Nejad, S. B., and Pittet, D., "Report on the Burden of Endemic Health Care-Associated Infection Worldwide.," *WHO Library Cataloguing-in-Publication Data* , 40 (2011).

[2] Haas, J. P. and Larson, E. L., "Measurement of compliance with hand hygiene," *Journal of Hospital Infection* **66**(1), 6–14 (2007).

[3] "SureWash." http://surewash.com. Accessed: 2017-08-08.

[4] Zhong, H., Kanhere, S. S., and Chou, C. T., "WashInDepth," *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services - MOBIQUITOUS 2016* **13**, 28–37 (2016).

[5] Xiang, L., Echtler, F., Kerl, C., Wiedemeyer, T., Lars, Hanyazou, Gordon, R., Facioni, F., Laborer2008, Wareham, R., Goldhoorn, M., Alberth, Gaborpapp, Fuchs, S., Jmtatsch, Blake, J., Federico, Jungkurth, H., Mingze, Y., Vinouz, Coleman, D., Burns, B., Rawat, R., Mokhov, S., Reynolds, P., Viau, P., Fraissinet-Tachet, M., Ludique, Billingham, J., and Alistair, "libfreenect2: Release 0.2," (2016). https://zenodo.org/record/50641.

[6] Chollet, F., "Keras." https://github.com/keras-team/keras (2015).

[7] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9351**, 234–241 (2015).

[8] Clevert, D.-A., Unterthiner, T., and Hochreiter, S., "Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs)," (2015).

[9] Kingma, D. P. and Ba, J. L., "Adam: a Method for Stochastic Optimization," (2015).

[10] Zeiler, M. D. and Fergus, R., "Visualizing and Understanding Convolutional Networks," *Computer VisionECCV 2014* **8689**, 818–833 (2014).