Hand Hygiene Poses Recognition with RGB-D Videos

Baiqiang Xia†, Rozenn Dahyot†, Jonathan Ruttle‡, Darren Caulfield‡ and Gerard Lacey†‡

† : *School of Computer Science and Statistics, Trinity College Dublin, Ireland*
‡ : *Glanta ltd, Surewash*

Abstract

Hand hygiene is the most effective way in preventing the health care-associated infection. In this work, we propose to investigate the automatic recognition of the hand hygiene poses with RGB-D videos. Different classifiers are experimented with the Histogram of Oriented Gradient (HOG) features extracted from the hand regions. With a frame-level classification rate of more than 95%, and with 100% video-level classification rate, we demonstrate the effectiveness of our method for recognizing these hand hygiene poses. Also, we demonstrate that using the temporal information, and combining the color with depth information can improve the recognition accuracy.

Keywords: Hand Hygiene, Poses Recognition, RGB-D

1 Introduction

According to the World Health Organization (WHO), hands are the main pathways of germ transmission in health care-associated infections (HCAI), which causes thousands of people deaths and billions of money losses each year. Hand hygiene plays a crucial role in the prevention of HCAI, which is usually achieved by rubbing hands with an alcohol-based formulation. In the WHO Guideline book of hand hygiene [WHO, 2009], routine gestures for hand hygiene with alcohol-based hand-rub formulation or soap[1] have been suggested. More specifically, some hand gestures of interest in the hand hygiene procedure are illustrated in Figure 1.

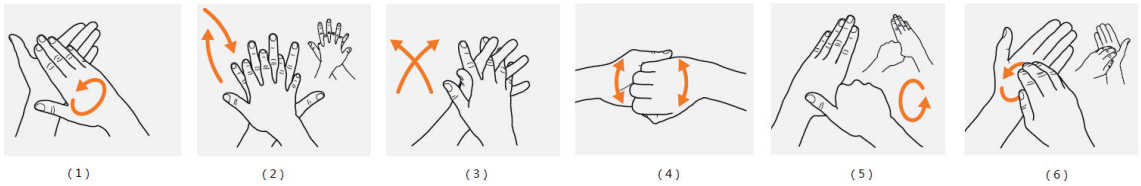

Figure 1: Hand hygiene Poses suggested by WHO. (1) : Rub hands palm to palm, (2) : Right palm over left dorsum with interlaced fingers and vice versa, (3) : Palm to palm with fingers interlaced, (4) : Backs of fingers to opposing palms with fingers interlocked, (5) : Rotational rubbing of left thumb clasped in right palm and vice versa, (6) : Rotational rubbing with clasped fingers of right hand in left palm and vice versa.

In the following, we first briefly review the vision based approaches for hand pose recognition with highlights on hand-washing poses recognition (section 2), and then present our own approach (section 3). Experimental results (section 4) show the effectiveness of the proposed approach in this hand hygiene recognition task, as well as the benefit of combining depth and color information, and using the temporal information. Finally, section 5 draws some conclusions of this work.

[1] Soaps are suggested when the hands are visibly soiled. If not, the alcohol-based hand-rub formulation is suggested to be used

 ISBN 978-0-9934207-0-2

2 State of the Art

Hand pose has attracted active research in computer vision domain, either for the hand pose estimation task which aims at the reconstruction of hand posture [Supancic III et al., 2015], or for the hand pose classification task which aims at recognizing a set of predefined hand gestures [Suarez and Murphy, 2012]. Concerning hand pose classification, it usually involves an early stage of hand localization in the video frames, followed by a machine learning stage which applies classifiers on extracted features to finally predict the pose [Suarez and Murphy, 2012, Sarkar et al., 2013]. For hand localization, various hand segmentation methods were proposed in consideration of the skin-color maps [Ibraheem et al., 2013]. These methods are claimed to suffer greatly from light changes, even with an illumination-invariant color schemes. In recent years, with the spread of commodity depth cameras, more and more works use simply a depth threshold to isolate the hands [Suarez and Murphy, 2012]. This type of method usually works with the presence of a single hand, which is also required to be the closest object to the camera. Ghobadi et al. [Ghobadi et al., 2007] explored the combination of the color and depth information for hand segmentation, in a pixel level clustering scheme. In the following machine learning stage, the hands are first represented by relevant features, such as color features [Llorca et al., 2011], shape features [Keskin et al., 2012, Suryanarayan et al., 2010], volume features [Suryanarayan et al., 2010], and temporal motion features [Elmezain et al., 2009]. The features are then fed to classifiers to predict the pose label, such as the Hidden Markov Model (HMM) [Kurakin et al., 2012], the Neural Networks [Hasan and Abdul-Kareem, 2014], the Support Vector Machine [Llorca et al., 2011], the Random Forest [Keskin et al., 2012], and the Linear Discriminant Analysis (LDA) [Wang and Zhang, 2013], etc.

Many applications have been derived from the related research, which mainly fall in the areas of human computer interaction (HCI) [Rautaray and Agrawal, 2015], robot control [Khan and Ibraheem, 2012] and human sign language recognition [Slama et al., 2014]. Concerning hand washing poses recognition, only a few works have been issued with videos from standard RGB cameras. In [Llorca et al., 2007, Llorca et al., 2011], researchers proposed to use a set of 21 binary SVM classifiers for recognizing the 6 hand washing gestures suggested in [RCN, 2005], together with another hand pose class defined as not belonging to any of the 6 poses. Prior statistics of skin/non-skin color, and motion information between frames are required in their method for hand segmentation. In both works, the videos were manually filtered and labeled by human experts on each frame. In [Hoey et al., 2010], a vision based system is proposed to assist people with difficulties in washing their hands. As far as we know, the depth information has not been investigated yet for this specific task.

3 Hand Hygiene Poses Recognition

We propose a computer vision system for recognizing these poses in RGB-D video streams of hands (cf. Figure 2). With a set of RGB-D hand hygiene videos, we first perform hand region segmentation on each frame with the Expectation Maximization technique working on the Gaussian Mixture Model. Then we extract the Histogram of Oriented Gradient (HOG) features on the hand regions, which are later processed with the Principal Component Analysis (PCA) for dimensionality reduction. The PCA transformed features are then used for training and testing, with different classifiers. Section 3.1 presents our strategy for hand segmentation. Features and classifiers are presented in section 3.2.

3.1 EM-based hands segmentation

To analyze the behavior of the hands in the videos, we are required to separate the hands from the background. To this end, for each pair of RGB and depth images, we first transformed the RGB image into the HSV color space, then constructed a feature vector (Fp) on each valid pixel using the Hue (h) and Saturation (s) information, together with the depth (d) information from the depth image. Formally, it means Fp={h,s,d}, with h, s, and d as integers in the scale of [0,255]. With these pixel-level features, we explored the Expectation Maximization (EM) technique with the Gaussian Mixture Models representation of the image features for hand segmentation. The Gaussian Mixture represents each segment of the features by a parametric Gaussian Distri-

 ISBN 978-0-9934207-0-2

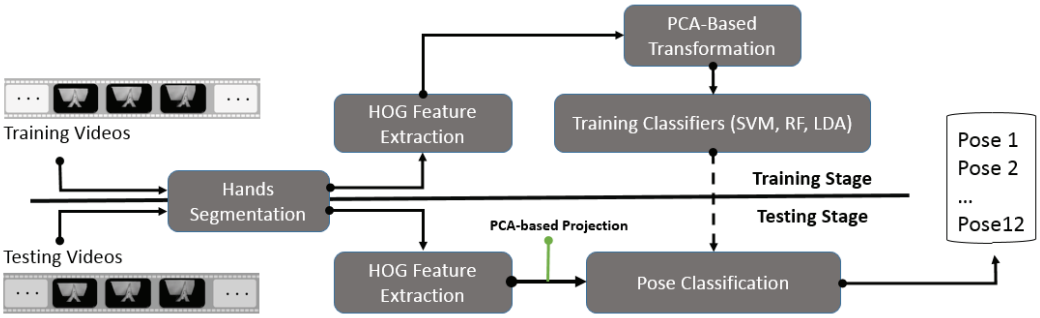

Figure 2: Overview of the proposed hand hygiene poses recognition pipeline.

bution. The Expectation Maximization (EM) technique then searches for the best parameters of the Gaussian models, which produce a maximum likelihood estimation with the given features. In the output of EM, each pixel is labeled with the Gaussian model with the highest probability to produce this pixel. It has been reported in [Ghobadi et al., 2007] that effective hand segmentation can be achieved using the EM technique and the Gaussian Mixture Models, based on the depth and color information. In our work, We find 4 Gaussian kernels provide the most accurate segmentation of hands in visual observation.

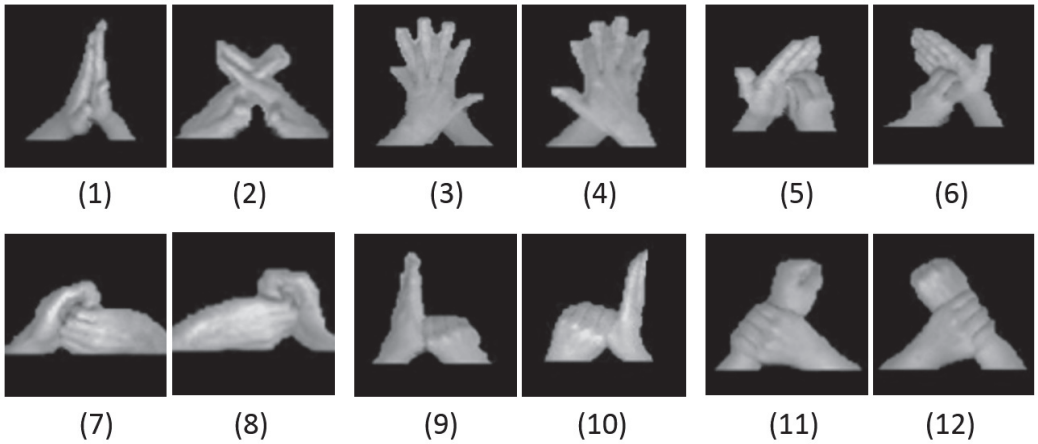

Figure 3: Illustration of segmentation output for 12 hand hygiene poses. (1) : Rub hands palm to palm; (2) : Rub interlaced fingers; (3) and (4) : Left palm over right dorsum with interlaced fingers and vice-versa; (5) and (6) : Rotational rubbing, backwards and forwards with clasped fingers of right hand in left palm and vice versa; (7) and (8) : Backs of right fingers to left palm with fingers interlocked and vice versa; (9) and (10) : Rotational rubbing of left thumb clasped in right palm and vice versa; (11) and (12) : Rotational rubbing left wrist clasped in right hand and vice versa.

3.2 Features Extraction and Machine Learning

With the segmented frames, we cut out the hand region, by removing the pixels below the row where the two arms are joined. The remaining part is then re-sized to a unified resolution of 80×80, for extraction of the Histogram of Oriented Gradient (HOG) features [Dalal and Triggs, 2005]. The HOG features are later explored in pose classification with different classifiers, namely the linear-kernel SVM classifier [Chang and Lin, 2011], the Linear Discriminant Analysis (LDA) classifier [Scholkopft and Mullert, 1999], and the Random Forest classifier [Breiman, 2001].

 ISBN 978-0-9934207-0-2

4 Experimental Results

4.1 Video Dataset and experimental design

We collected 72 RGB videos and 72 corresponding depth videos from 6 different subjects using the SoftKinetic DS325 camera. Each subject performs 12 different hand hygiene poses above a planar table, with the camera projecting vertically downwards. Each video lasts 2-4 seconds. With a frame rate of 25 per second, it results in 42-110 frames in each video. As there is misalignment between the images from RGB and depth sensors due to different sensor positions, we warped each color frame to the corresponding depth frame with the warping matrix provided by the camera system. In Figure 4, we illustrate an episode of a depth video and the corresponding RGB color video after warping.

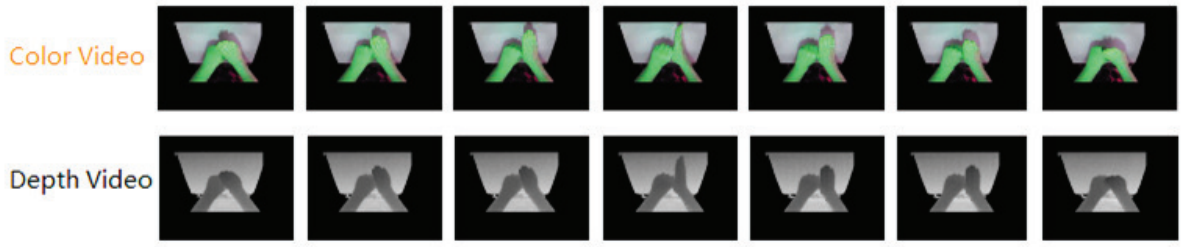

Figure 4: Example of an RGB Video and the corresponding Depth Video

We use the Leave-One-Person-Out (LOPO) subject-independent cross-validation protocol in evaluating the performance of the pose classification method, where each subject is used for the testing step once, with the remaining subjects used in the training step. This protocol enables the largest number of instances in training step, with the requirement that no subjects should enroll in both training and testing steps at the same round. Under this protocol, we explore the usage of the RGB channel (using the corresponding grayscale values), and the depth channel separately, in hand hygiene poses recognition. To address the high dimensionality of the HOG features (2916 dimensions), in each round of the cross-validation, the original HOG features are projected in lower dimensional subspace using the Principal Component Analysis (PCA). We note here that, in each round, the projection matrix of PCA is learned on the training set alone.

4.2 Frame-level Pose recognition

Figure 5 shows the pose classification results when recognizing on each frame of the videos. The x-axis shows the dimensionality of the PCA-based feature subspace, and the y-axis shows the average recognition rate over all the frames, for different classifiers. In Figure 5, for both the color and depth channels, the LDA classifier outperforms the Linear-SVM classifier, and further outperforms the Random Forest classifier. In Figure 5 (a), with 75 dimensions of the PCA-based features, the the LDA classifier achieves 94.80% pose recognition rate. The corresponding results for the Linear-SVM is 93.17%, and is 91.87% for Random Forest. With the depth information, as shown in Figure 5 (b), the LDA classifier achieves 92.35% classification rate using 100 dimensions of the PCA-based features. Correspondingly, the Linear-SVM classifier achieves 89.75%, and the Random Forest classifier reaches 87.97%. These results demonstrate the effectiveness of our hand hygiene pose recognition approach, as well as the stability of the PCA projections in different rounds of the cross-validation. Additionally, the color channel generally outperforms the Depth channel. We assume that, apart from the data modality differences, the low quality of the depth data provided by the camera also accounts for this. In Table 1, we show the average time consumption for classifying each frame, concerning the above results, generated by an Intel Core i7 CPU 3.70 GHZ with 16GB of RAM, in Matlab implementation. Again, the LDA classifier demonstrates significant merit over the Linear-SVM. Notably, the Random Forest classifier results in the smallest time consumption. Considering both the recognition rate and time consumption, **we choose the results from the LDA classifier for further analysis**.

 ISBN 978-0-9934207-0-2

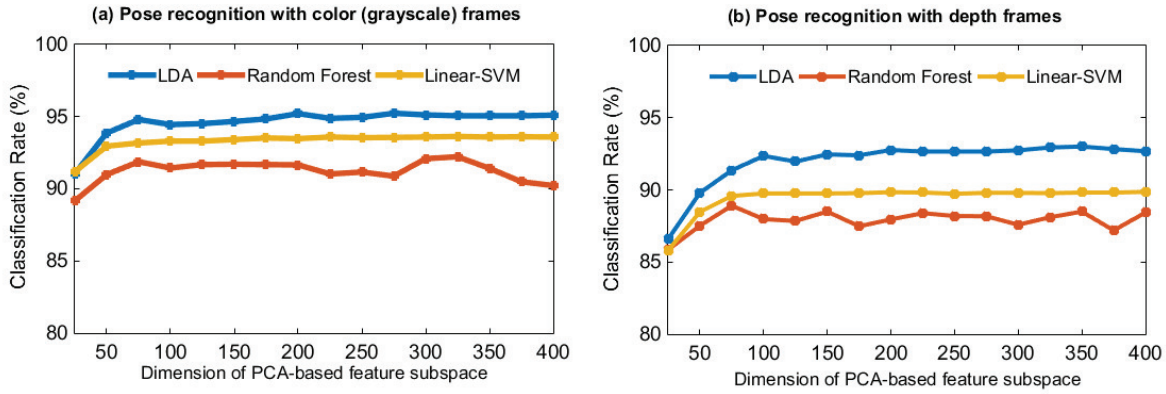

Figure 5: Frame by Frame Pose recognition Results

	LDA	Random Forest	Linear-SVM
Color	0.0100 ms	0.0077 ms	0.6774 ms
Depth	0.0139 ms	0.0086 ms	0.8372 ms

Table 1: Time Consumption for classifying each frame (millisecond).

Apart from the above, another perspective of the results is the confusion matrix, which shows the interaction of different classes during classification. Thus, we show in Table 2 and Table 3 the confusion matrices of the recognition results from the LDA classifier, with the color frames and with the depth frames, respectively. In the two tables, the ground truth goes with the row labels, and the pose estimation result goes with the column labels. In both confusion matrices, the diagonal parts have the dominant accumulation of the data. Very few instances are confused, as shown in the off-diagonal parts of the matrices (blanks mean 0). It means that our pose recognition approach is not biased to particular poses.

The closest work to ours in the state of the art is presented in [Llorca et al., 2011], in which the researchers performed the recognition of 7 types of hand poses with the RGB videos. With manually filtered and labeled frames in 6 videos for training and frames of another 2 videos used for testing, they achieved 82.29% correctness over the 7 pose classes, using 21 binary SVM classifiers. No cross-validation scheme was issued in their work. In comparison, as shown in Figure 5 (a) and Table 2, we have achieved a more promising classification rate of 94.80% over 12 hand pose classes in subject-independent cross-validation, with a single LDA classifier which also works much faster than SVM as shown in Table 1. In addition, Figure 5 (b) and Table 3 also make the first published results for hand hygiene gesture recognition using the depth information.

4.3 Video-level Pose recognition results

Although the frame-level pose recognition has demonstrated its effectiveness, the temporal evolution information of the video frames has been omitted. We are motivated to explore the usage of the temporal relationship of video frames. With this concern, we switch from recognizing the pose in the frames, to recognizing the pose in the episodes. By applying the sliding-window technique on the temporal frames, we formatted a set of video episodes. The poses in these episodes are recognized as the majority voting results of the results from the containing frames. As the window size controls the length of the episode (number of frames), it further influences the recognition accuracy. Thus, we show in Figure 6 the relationship between the window size and the recognition rate. To combine the contributions from both channels, we also propose a fusion scheme which performs majority voting on frames of both the color and the depth channels within each episode. For the

 ISBN 978-0-9934207-0-2

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12
G1	315	1			2				1	4	6	
G2		308									1	
G3		8	377	8								
G4			1	395								
G5	2	5			337	2	4			12	22	
G6						406				8	6	
G7							434			31	1	
G8					4			349		14		
G9	10	1							382		31	
G10		2			7		2	1		369	9	1
G11	1	4				1	2		1		379	
G12		3									21	324

Table 2: Confusion matrix of pose recognition with color frames. G: Ground truth, E: Estimated label

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12
G1	307				1						21	
G2		307									2	
G3		2	362	26								3
G4		1		394	1							
G5	10	2	2		316		7	12	2	20	4	9
G6						413	1			3	3	
G7	1						433			17	4	11
G8					19			300		47	1	
G9	13	4		2					385		20	
G10	1	1			5		8	9	3	356	4	4
G11	26	1				1	13				347	
G12										3	3	342

Table 3: Confusion matrix of pose recognition with depth frames. G: Ground truth, E: Estimated label

majority voting procedures, we use the same frame-level results as for the previous confusion matrices.

In Figure 6, we observe that, for both the color and the depth channels, the larger the sliding window size, the higher the pose recognition rate. With a window size of 10, we achieve 98.99% and 97.86% recognition rate for color and depth channels, respectively. With window size growing to 20, the corresponding recognition rates reach 99.38% and 98.31%. These results outperform significantly the corresponding frame-level results, which were 94.80% in the color channel and 92.35% for the depth channel. Apart from this, when we take each video as one episode (not shown in the Figure), both the color and depth channels achieve 100% classification rate. These results demonstrate that the temporal information gives enhancement to this pose recognition problem. Also, the fusion scheme obviously outperforms each single channel. With window size of 10, we achieve 99.17% pose classification rate. It demonstrates that the combination of the color and the depth channels also enhances the pose recognition performance.

5 Conclusion

In this work, we proposed a computer vision framework to recognize the hand hygiene poses from RGB-D videos. Using the LDA classifier on PCA-projected HOG features, we achieve >95% frame-level recognition

 ISBN 978-0-9934207-0-2

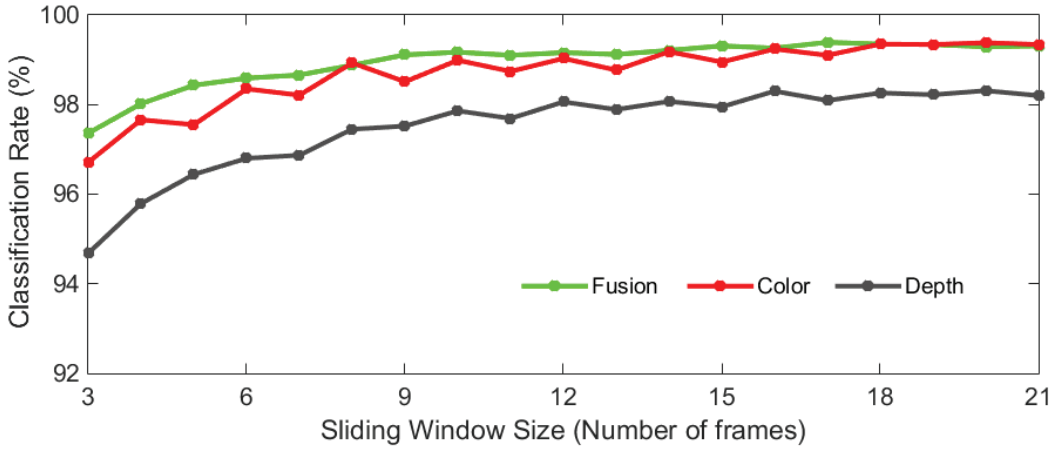

Figure 6: Pose recognition rates with different sliding window size

rate, and 100% video-level classification rate. It demonstrates the capability of the proposed method in discriminating the hand-hygiene poses. In addition, the temporal information further improves the performance of our system. We also show that the fusion of color and depth channels is beneficial to the recognition performance. The resulted system could be applied in hand hygiene monitoring, education, and also could be extended to more constrained scenarios, such as in surgery preparation where additional hand hygiene poses are required.

Acknowledgments

This work has been supported by the Innovation partnership project (IP-2014-0290) funded by Enterprise Ireland, the European Regional Development Fund, Movidius.com and SureWash.com.

References

[Breiman, 2001] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

[Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

[Dalal and Triggs, 2005] Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE.

[Elmezain et al., 2009] Elmezain, M., Al-Hamadi, A., Pathan, S. S., and Michaelis, B. (2009). Spatio-temporal feature extraction-based hand gesture recognition for isolated american sign language and arabic numbers. In *Image and Signal Processing and Analysis, 2009. ISPA 2009. Proceedings of 6th International Symposium on*, pages 254–259. IEEE.

[Ghobadi et al., 2007] Ghobadi, S., Loepprich, O., Hartmann, K., and Loffeld, O. (2007). Hand segmentation using 2d/3d images. In *IVCNZ 2007 Conference, Hamilton, New Zealand*, volume 5.

[Hasan and Abdul-Kareem, 2014] Hasan, H. and Abdul-Kareem, S. (2014). Static hand gesture recognition using neural networks. *Artificial Intelligence Review*, 41(2):147–181.

[Hoey et al., 2010] Hoey, J., Poupart, P., von Bertoldi, A., Craig, T., Boutilier, C., and Mihailidis, A. (2010). Automated handwashing assistance for persons with dementia using video and a partially observable markov decision process. *Computer Vision and Image Understanding*, 114(5):503–519.

[Ibraheem et al., 2013] Ibraheem, N. A., Khan, R. Z., and Hasan, M. M. (2013). Comparative study of skin color based segmentation techniques. *International Journal of Applied Information Systems (IJAIS)*.

[Keskin et al., 2012] Keskin, C., Kıraç, F., Kara, Y. E., and Akarun, L. (2012). Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In *Computer Vision–ECCV 2012*, pages 852–863. Springer.

[Khan and Ibraheem, 2012] Khan, R. Z. and Ibraheem, N. A. (2012). hand gesture recognition: a literature review. *International Journal of Artificial Intelligence & Applications (IJAIA)*, 3(4).

[Kurakin et al., 2012] Kurakin, A., Zhang, Z., and Liu, Z. (2012). A real time system for dynamic hand gesture recognition with a depth sensor. In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 1975–1979. IEEE.

[Llorca et al., 2007] Llorca, D., Vilarino, F., Zhou, Z., and Lacey, G. (2007). A multi-class svm classifier ensemble for automatic hand washing quality assessment. In *BMVC Proc. Brit Mach Vision Conference, Warwick, UK*, pages 213–223.

[Llorca et al., 2011] Llorca, D. F., Parra, I., Sotelo, M. Á., and Lacey, G. (2011). A vision-based system for automatic hand washing quality assessment. *Machine Vision and Applications*, 22(2):219–234.

[Rautaray and Agrawal, 2015] Rautaray, S. S. and Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43(1):1–54.

[RCN, 2005] RCN (2005). Methicillin resistant staphylococcus aureus (mrsa): guidance for nursing staff. *http://www.nhs.uk/conditions/mrsa/documents/*.

[Sarkar et al., 2013] Sarkar, A. R., Sanyal, G., and Majumder, S. (2013). Hand gesture recognition systems: a survey. *International Journal of Computer Applications (0975–8887)*, 71(15).

[Scholkopft and Mullert, 1999] Scholkopft, B. and Mullert, K.-R. (1999). Fisher discriminant analysis with kernels. In *Proceedings of the 1999 IEEE Signal Processing Society Workshop Neural Networks for Signal Processing IX, Madison, WI, USA*, pages 23–25.

[Slama et al., 2014] Slama, R., Wannous, H., and Daoudi, M. (2014). Grassmannian representation of motion depth for 3d human gesture and action recognition. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 3499–3504. IEEE.

[Suarez and Murphy, 2012] Suarez, J. and Murphy, R. (2012). Hand gesture recognition with depth images: A review. In *RO-MAN, 2012 IEEE*, pages 411–417.

[Supancic III et al., 2015] Supancic III, J. S., Rogez, G., Yang, Y., Shotton, J., and Ramanan, D. (2015). Depth-based hand pose estimation: methods, data, and challenges. *arXiv preprint arXiv:1504.06378*.

[Suryanarayan et al., 2010] Suryanarayan, P., Subramanian, A., and Mandalapu, D. (2010). Dynamic hand pose recognition using depth data. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 3105–3108. IEEE.

[Wang and Zhang, 2013] Wang, Y. and Zhang, L. (2013). 3d hand gesture recognition based on polar rotation feature and linear discriminant analysis. In *Intelligent Control and Information Processing (ICICIP), 2013 Fourth International Conference on*, pages 215–219. IEEE.

[WHO, 2009] WHO (2009). WHO guidelines on hand hygiene in health care. *Accessed via http://www.who.int/gpsc/5may/tools/9789241597906/en/*.