A multi-class SVM classifier ensemble for automatic hand washing quality assessment

D.F. Llorca, F. Vilarino, Z. Zhou and G. Lacey
Graphics, Vision and Visualisation Group (GV2)
Computer Science Dep. Trinity College Dublin, Rep. of Ireland
llorca@depeca.uah.es

Abstract

Hand washing is a critical activity in preventing the spread of infection in health-care environments. Several guidelines recommended a hand washing protocol consisting of six steps that ensure that all areas of the hands are thoroughly cleaned. In this paper, we describe a novel approach that uses a computer vision system to monitor the user's hands motions in order to ensure that the hand washing guidelines are followed. This work presents two main contributions: a description of a system which delivers robust segmentation of the hands using a combination of colour and motion analysis, and the implementation of a multi-class classification of the hand gestures using a SVM ensemble. The system performance is analysed and compared with human performance, showing an accuracy close to that of human experts.

1 Introduction

Hand washing is widely acknowledged to be the single most important activity for reducing the spread of infection in clinical environments. It is extremely important that health care professionals use the correct technique. This means that no areas of the hands should be missed. According to several guidelines [15] the correct hand washing procedure should comprise six different poses as depicted in Figure 1:

Pose 1 Pose 2 Pose 3 Pose 4 Pose 5 Pose 6

Figure 1: *Pose 1*: Rub palm to palm. *Pose 2*: Rub palm over the back of the other one. *Pose 3*: One hand over back of the other hand and rub fingers. *Pose 4*: Rub palm to palm with fingers interlaced. *Pose 5*: Wash thumbs of each hand separately using rotating movement. *Pose 6*: Rub finger tips against the opposite palm using circular motion.

In the last few years, several authors have addressed the problem of single hand gesture recognition [16], and recently gesture recognition has been applied to hand washing [7]. Among the wide range of vision-based hand gesture recognition techniques in the literature, we emphasize: Histograms of oriented gradients (HOG) have been used as feature

BMVC 2007 doi:10.5244/C.21.40

vectors in hand gesture classification [5]. Spatio-temporal hand gesture recognition using neural networks has been introduced in [19] and [12], and recognition of gestures using hidden Markov models (HMM) has been also widely studied [16],[18], [14]. In [10], a robust hand posture detection method based on the Viola and Jones detector [20] is proposed; in addition, a frequency analysis-based method for class separability estimation is used and a comparison of non-cascaded and cascaded detectors is presented.

In this work, a novel system for hand washing quality assessment using computer vision is proposed. Hands/arms segmentation is achieved by combining skin and motion features to ensure a robust region of interest (ROI) selection process which is independent of lighting conditions or reflectivity of the sink (ceramic, stainless steel, etc.). Distributed HOG are used to create the feature vector which will be dispatched to the classifier. In order to recognize the six different poses of the correct hand washing technique, a support vector machine (SVM) [1] with a one-against-one multi-class approach is used.

This paper is organised as follows: Section 2 provides a detailed description of the ROI selection process. The feature extraction method is described in Section 3, and the description of the classifier used to recognise the different poses is shown in Section 4. The results achieved up to date and their comparison with human performance are presented in Section 5. Finally, our conclusions and the description of our future lines of research are presented in Section 6.

2 ROI Selection

To increase the reliability of the classification it is necessary to identify the hands in the images and to locate a region of interest around the hands. This is important for both the case when the hands are joined and when they are separate. This step is crucial, since both the training and performance of the classifier depend on the manner in which the ROI is calculated.

Typically in the case of an acquisition process obtained from a stationary camera, background subtraction would be a good segmentation strategy. However, in this application the foreground drastically changes the background, not only its intensity but also its colour. With fixed lighting conditions in stainless steel sinks, for example, the hands/arms may tend to be confused with background due to the multiple reflections and changes in colour on the steel surface. In addition, adaptive background subtraction approaches appear not to improve the results because almost all the time the hands are moving around the same area, and consequently the hands are gradually integrated with the background. Accordingly, other techniques should be used. Hands/arms segmentation is thus posed in this work as a problem which integrates both skin detection and motion analysis.

2.1 Skin Colour Detection

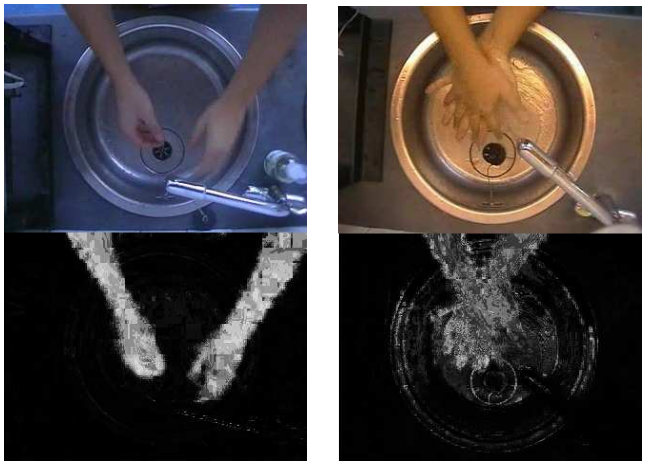
Skin detection plays an important role in several applications such as face detection, searching and filtering image content on the web, video segmentation, face/head tracking, etc. In our case a non-parametric skin modeling with Bayesian models [9] is applied, in which the skin and non-skin colours are modelled through histograms. We quantize the colour space *RGB* to a number of bins N^{bins} and count the number of colour pixels in each bin N_{skin} for skin class as well as $N_{non-skin}$ for non-skin class. Finally we normalise each bin to get the discrete conditional *skin/non-skin* colour distribution

$P(rgb|skin)/P(rgb|non\text{-}skin)$ [9]. A skin binary map can be made upon a properly selected threshold over the posterior probabilities obtained.

Although the described scheme can successfully deal with lighting changes, the varying lighting conditions of the wash hand basin requires a lightning compensation step in order to achieve more robust estimates of skin tone. We propose the use of the grey-world approach implemented by Chen et al. [6]. This method is based on the assumption that the spatial average of surface reflectance in a scene is achromatic. Since the light reflected from an achromatic surface is changed equally at all wavelengths, it follows that the spatial average of the light leaving the scene will be the colour of the incident illumination. In addition, even though lighting compensation reduces the variability in skin tones detected, there remains a significant challenge to robust segmentation in real world situations, i.e., reflections due to stainless steel wash hand basins. This is a factor of both the skin tones of the user and the lighting conditions, and an example of this can be observed in Figure 2. In Figure 2 (a) reliable segmentation is achieved, while in Figure 2 (b) additional light from a window produces certain reflections of the hands on the stainless steel. Consequently, in situations where stainless steel wash hand basins are used, additional features are required to achieve robust segmentation. In our approach we propose the use of features based on motion analysis.

2.2 Skin and Motion Segmentation

The idea on which our proposal for robust hand segmentation is based consists of computing an overall image that integrates both skin and motion features. For each skin pixel, which could be a false positive, we calculate the motion vector obtained by means of optical flow analysis [13], after applying an average filter over the motion magnitude image. If the averaged motion magnitude is greater than a fixed threshold T_{ID} the probability estimate is increased. If it is less than T_{ID}, then the probability is decreased. We implement



(a) (b)

Figure 2: Upper row: Original images. Lower row: Skin probability maps. (a) With favorable, and (b) unfavorable lighting conditions.

this increase/decrease effect by means of the coefficients I and D, defined according to the following equations:

$$I = I_F(1 - \frac{1}{\exp(\frac{M_i - T_{ID}}{0.5T_{ID}})}); \qquad D = D_F \frac{1}{\exp(\frac{M_i}{0.5T_{ID}})} \tag{1}$$

where I_F and D_F are the increase and decrease factors, M_i is the motion magnitude of pixel i, and T_{ID} is the threshold value which decides if there is going to be an increase or a decrease in the output image. The increase function should reach the maximum value quickly, whereas decrease function must be less steep. This is because we want to quickly highlight motion as soon as it is detected and penalise its absence more slowly. Thus skin pixels due to metal sinks and specular reflections are slowly removed while the hands/arms are correctly segmented only if they are moving. This architecture does not resolve the issue of hand detection while the hands are not moving. However, in this application the hands rarely stop moving and one can argue that if they do stop no effective hand washing is being performed. Figure 3 shows a summary of the global architecture described so far for the hands detection based on both colour and motion information.

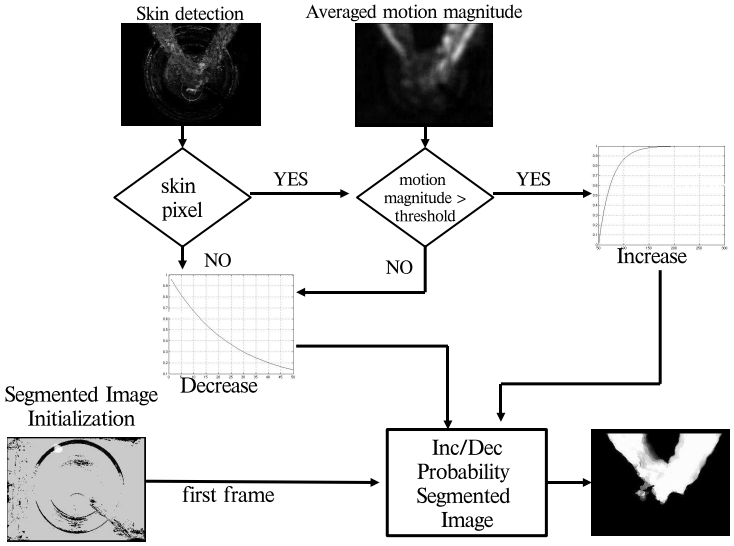


Figure 3: Skin and motion detector for hands/arms segmentation scheme

Following detection, we apply a connected components analysis to identify the hands and arms. Geometric analysis of the shapes leads to the computing of the location of the boundaries of the ROI. Next, a region of interest covering the hands is selected. The vertical symmetry axis is used for computing upper and lower boundaries, which are then used for obtaining the lateral ones. This process is depicted in Figure 4.

3 Feature Extraction

Choosing discriminating and independent features is key to any pattern recognition algorithm being successful in classification. We use local histograms of oriented gradients as

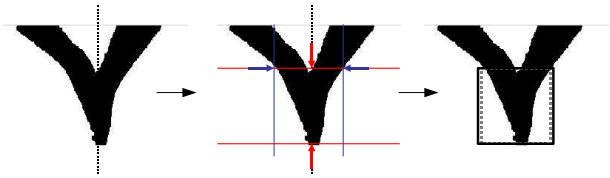


Figure 4: Hands ROI selection: First a vertical symmetry axis is obtained, then the vertical bounds are delimited, and finally the horizontal bounds are obtained.

our single frame feature extraction method, which has reported good performance results in recent applications [4]. The aim of this method is to describe an image by a set of local histograms. These histograms count occurrences of gradient orientation in a local part of the image. The procedure is a follows: First the colour image is converted to grey level, then the gradient is calculated, and next the image is split into cells which are defined as square regions with a predefined size in pixels. For each cell, we compute the histogram of gradients by accumulating votes into bins for each orientation. Votes can be weighted by the magnitude of the gradient vector, so that the histogram takes into account the importance of the gradient at a given point.

Due to the variability in the images, it is necessary to normalise the cell histograms. Cell histograms are locally normalised according to values of the neighbored cell histograms. The normalisation is done among a group of cells, which is referred to as a block. A normalisation factor is then computed over the block and all histograms within this block are normalised according to this normalisation factor. We use the L2-norm scheme:

$$v \rightarrow \frac{v}{\sqrt{\|v\|_2^2 + \varepsilon}} \tag{2}$$

where ε is a small regularization constant needed because sometimes empty gradients are going to be evaluated. Note that according to how each block has been built, a histogram from a given cell can be involved in several block normalisations. Thus we are going to have some redundant information which, according to the work of Dalal et al. [4], improves the performance. Figure 5 shows a graphical description of this method.

When all histograms have been computed for each cell, we build the description vector of an image by concatenating all histograms into a single vector. In order to compute the vector dimension several parameters have to be taken into account: ROI dimension, cell size, block size, number of bins and number of overlapped blocks. Joining together all the elements described so far, the final recount of features is a follows: the ROI size is 128×128, each window is divided into cells of size 16×16 and each group of 2×2 cells is integrated into a block in a sliding fashion, so blocks overlap with each other. Each cell consists of a 16-bin HOG and each block contains a concatenated vector of all its cells. Each block is thus represented by a 64 feature vector which is normalised to an L2 unit length. Thus, in our case we have feature vectors of 3.136 dimensions. The following chart summarizes the list of all the features included in the final feature vector:

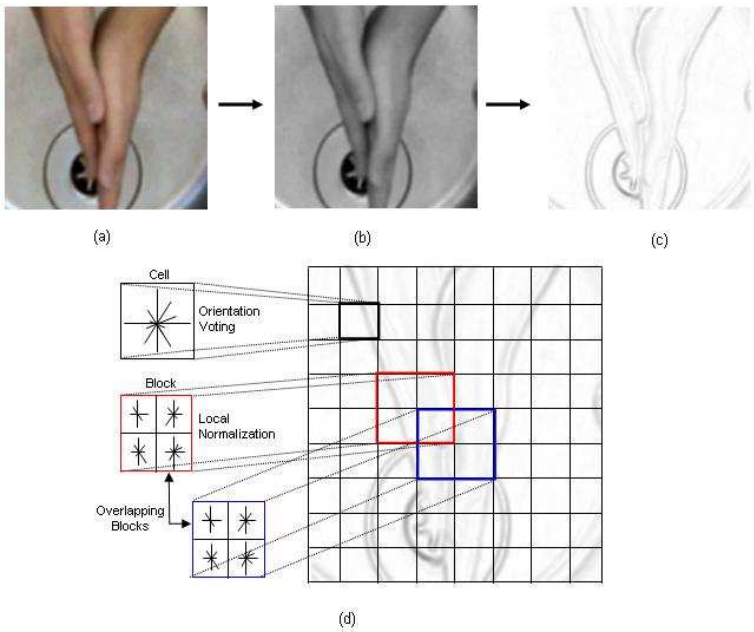


Figure 5: HOG description, (a) original image, (b) grey image, (c) gradient image, (d) HOG, cells distribution and blocks normalisation

$Cells_{Row} = 128/16 = 8$	$Cells_{Col} = 128/16 = 8$
$CellsBlock_{Row} = 32/16 = 2$	$CellsBlock_{Col} = 32/16 = 2$
$OverlapBlocks_{Row} = 8-2+1 = 7$	$OverlapBlocks_{Column} = 8-2+1 = 7$
$\mathbf{VectorDIM = 7*7*2*2*16 = 3136}$	

4 The SVM Classifier Ensemble

The support vector machine classifier is a binary classifier algorithm that looks for an optimal hyperplane as a decision function in a high dimensional space [1]. It is a kind of example-based machine learning method for both classification and regression problems. This technique has several features that make it particularly attractive. Traditional training techniques for classifiers, such as multi-layer perceptrons (MLP) use empirical risk minimisation and only guarantee minimum error over the training set. In contrast, SVM bases on the structural risk minimisation principle [1] which minimises a bound on the generalisation error and therefore should perform better on novel data.

In order to perform the multi-class classification the one-against-one method is proposed, with which $k(k-1)/2$ different binary classifiers are constructed and each one trains data from two different classes. In this application, there are 6 different classes relating to the 6 different poses described in Figure 1, plus an additional complementary class representing an indeterminate pose (with $k = 7$ we have 21 classifiers). In practice one-against-one method is one of the more suitable strategies for multi-classification problems [8]. After training data from the *i*th and the *j*th classes, the following voting strategy is used: if a binary classifier says that a given sample x is associated to class j,

then the vote for the class j is incremented by one. Finally, sample x is predicted to be in the class with the largest vote - this approach is referred as the *max-wins* strategy [21]-.

5 Results

We tested our system in a database of hand poses which was built up in the following way: We recorded 6 videos about 600 frames each (24 seconds long). Each video consisted of a sequence of hands performing the movements associated to only one specific pose. Next, the experts labelled each frame as pose 1 to 6, using the label 7 for the indeterminate pose, and those frames in which the experts did not agree were rejected. Following, the selected frames were analyzed by the ROI segmentation procedure described in Section 2, and all those frames in which the ROI could not be detected by the system because the hands were not joined -basically at the start and at the end of the video sequences- were rejected. The remaining frames -about a half of the original data set- constitute the training set with which the SVM ensemble was trained. The test set consisted of two videos of about 1 minute long in which the different poses were randomly visualized. The 3 experts performed a labelling of all the frames for these 2 videos, and all those frames in which the experts did not agree were rejected. In addition, all the frames in which a ROI was not segmented by the system were not taken into account. The remaining frames (1449/1924 for video 1 and 1025/1309 for video 2) constitute the test set.

Skin and non-skin reference histograms were obtained from the open source filtering Poesia project [17]. The project compiled 32^3 3D RGB histograms from the Compaq database and left them available in two files, one for skin pixels and other for non-skin pixels. Prior probabilities of skin and non-skin were assumed to be 0.4 and 0.6, which is a fair approach of the average region covered by the arms within the region of analysis. The segmented ROIs, from where features were extracted, were resized to a fixed size of 128×128 in order to have a normalised area of analysis. For the motion analysis we applied the Lucas and Kanade approach [13] with a threshold value $T_{ID} = 50$ and a window of 7 pixels for the average filter. We used the LIBSVM library [3] for building up the SVM classifier ensemble with radial basis function (RBF) kernels. All the single classifiers used a $\gamma = 3.2e^{-04}$ as the RBF parameter, which was obtained empirically.

We can see the results after applying multi-classification with HOG features and SVM classifier in Table 1. Although detection rate is low for the "other poses" case, which is reasonable taking into account that it is the class with higher variability, all the recommended poses are classified with detection rates greater than 85%. Three of them are classified with an accuracy greater than 91%, and the best classified class has a detection rate of 96.09%. The last row of Table 1 shows the inter-observer agreement as the percentage of coincident labels assigned by different experts to frames of the same class, which provides information about the intrinsic difficulty of each pose. The inter-observer agreement was calculated in the following way: We take all the M^1 frames labelled as class 1 by Expert 1 and Expert 2 and we count the number of coincidences $N_1^{1,2}$. The inter-observer agreement for Experts 1 and 2 in class 1 is calculated as the ratio $N_1^{1,2}/M^1$. We repeat the same procedure for Experts 2 and 3, and for Experts 1 and 3, and we average the results (this same scheme is applied for the rest of the classes).

The results shown in Table 1 appear to confirm that the proposed approach behaves in a similar way as the experts: Those classes which show the highest detection rate, as in the

Data sets	**Pose 1**	**Pose 2**	**Pose 3**	**Pose 4**	**Pose 5**	**Pose 6**	**Others**
Training Data Set	638	780	594	977	902	1073	587
Test Data Set	488	414	360	456	477	358	304
Detection rate	**86.07%**	**91.55%**	**94.72%**	**89.25%**	**86.37%**	**96.09%**	**61.84%**
Inter-observer agr.	94.59%	97.71%	97.20%	97.53%	98.46%	97.43%	82.98%

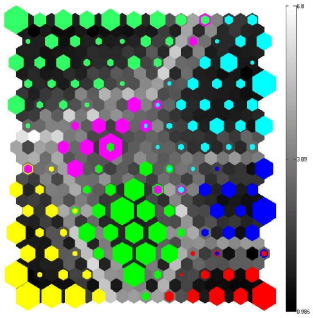
Table 1: Detection rate for the different classes

case of *Pose* 6, tend to show a high expert agreement, and viceversa -both low agreement and detection rate in *Pose* 7-, perhaps with an exception in the case of *Pose* 5. This fact can be analysed in the distribution table of Figure 6 (a), in which the element in the row i and column j represents the number of frames labelled by the human as *Pose* i which were classified by the system as *Pose* j. The distribution matrix shows how most of the misclassified samples of *Pose* 5 are voted for *Pose* 6 and the "other poses" class. However, the system appears to achieve a better generalization of *Pose* 6, since no misclassified frames are voted for *Pose* 5 in this case. Figure 6 (b) shows the cluster representation obtained using a self organizing map (SOM) [11] after a principal component analysis (PCA) holding 95% of the eigenvector spectrum, which produces a 2D mapping from the original feature space suitable for visualization purposes. Each sample in the data set is associated to one cell. The coloured cells in the SOM represents hit histograms of the different classes -the size of each cell being related to the number of samples in the cluster-, while the grey scale represents distances between cells. The SOM shows a clear cluster distribution, with several multi-class cells in the boundaries between clusters. This fact make us think that a simpler classification strategy specifically focusing on these regions could improve the classification results.

Figure 7 depicts an example of the single frame classification results in a given sequence. The detected class is shown with a number at the top left of each image. The system correctly detects when the hands are separate or joined. Even with a single frame approach the classifier correctly detects the transitions between poses, which are classified as the "other poses" class (a video sample of the system output can be obtained via web in the following link: http://www.cs.tcd.ie/VAMP/handwashing.avi).

Poses		SYSTEM						
		1	2	3	4	5	6	7
HUMAN	1	240	0	2	8	0	0	4
	2	0	198	0	0	14	0	0
	3	0	3	170	0	2	0	0
	4	1	0	0	292	2	4	0
	5	3	0	0	3	194	11	5
	6	3	0	0	0	0	143	2
	7	1	7	0	5	18	5	109

(a)



(b)

Figure 6: (a) Distribution table of the multi-class ensemble. (b) Cluster representation using SOMs. The grey scale represents distance between cells and each color is associated with a class-cluster. The size of each cell is related to the number of samples in the cluster.

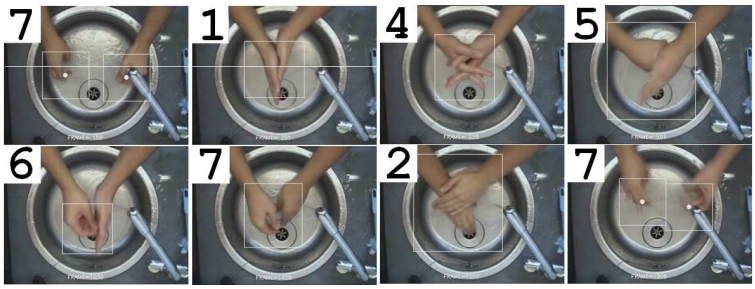


Figure 7: A sequence of frames with their associated pose class.

6 Conclusions and Future Work

In this work we presented a hand washing quality assessment system which can be used in hospitals, clinics, food processing industry, etc. One camera is placed over the sink focusing on the hand washing area. The system is implemented in a PC-based architecture and the complete algorithm runs at an average rate of 20 frames/sec. Hands/arms are segmented by means of skin and motion features. This segmentation process is robust against specular reflections due to steel sinks, illumination changes, etc. A SVM multi-class classification scheme using 21 binary classifiers, combined with histograms of oriented gradients locally distributed and normalised features, appear to yield very acceptable results. Hand washing poses that can be either done with left or right hand (symmetric poses) are correctly detected. The worst detected class is the one with higher variability (*other poses*) which has a detection rate of 61.84%. The minimum and maximum detection rate for the six poses are 86.07% (*Pose* 1) and 96.09% (*Pose* 6).

As a novel application, the outcomes shown in this contribution are encouraging results. However, several tasks are planned for the future: Hands/arms segmentation will be improved to detect the hands even if they are not moving. The search of an optimal tradeoff between the discrimination power of the feature vector and its size is an important issue in order to reduce the overall computation load in on-line systems. Single frame classification can be improved by adding a multi-frame validation process, in which stochastic models (hidden Markov models) could play an important role integrating sequence information. The use of additional features, particulary those based on motion by means of optical flow analysis and motion history gradients [2], are being investigated. Finally, although the multi-class SVM ensemble appears to show good classification results, the use of other alternatives with a lower computational cost are in our scope, looking forward to integrate this application into a feasible low cost FPGA platform.

References

[1] B.E Boser, I.M Guyon, and V.N Vapnik. A training algorithm for optimal margin classifiers. In *ACM Workshop on COLT*, pages 144–152, 1992.

[2] G.R. Bradsky and J.W. Davis. Motion segmentation and pose recognition with motion history gradients. *Machine Vision and Applications*, 13:174–178, 2002.

[3] C.C Chang and C.J Lin. LIBSVM: A library for support vector machines. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[4] N Dalal and B Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, volume 2, pages 886–893, 2005.

[5] W. T. Freeman and M. Roth. Hand gesture recognition using orientation histograms. In *Proc. IEEE TENCON*, pages 1355–1368, 1999.

[6] B. Funt, K. Bernard, and L. Martin. A fast skin region detector. In *ESC Division Research*, 2005.

[7] J. Hoey, A. von Bertoldi, et al. Assisting persons with dementia during handwashing using a partially observable Markov decision process. In *Proc. ICVS*, 2007.

[8] C.W Hsu and C.J Lin. A comparison of methods for multi-class support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425, 2002.

[9] M.J Jones and J.M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46(1):81–96, 2002.

[10] M. Klsch and M. Turk. Robust hand detection. In *Proc. IEEE CAFGR*, pages 614–619, 2004.

[11] T. Kohonen. *Self-Organized Maps*. Springer, Heidelberg Berlin, 1995.

[12] D. Lin. Spatio-temporal hand gesture recognition using neural networks. In *Proc. IEEE WCCI*, volume 3, pages 1794–1798, 1998.

[13] B.D. Lucas and Kanade T. An iterative image registration technique with an application to stereo vision. In *Proc. IUW*, pages 121–130, 1981.

[14] S. Marcel, O. Bernier, et al. Hand gesture recognition using input-output hidden markov models. In *Proc. IEEE CAFGR*, pages 456–461, 2000.

[15] MRSA. Methicillin-resistant staphylococcus aureus. Guidance for nursing staff. In *Royal College of Nursing*, 2005.

[16] V. Pavlovic, R. Sharma, and T Huang. Visual interpretation of hand gestures for human-computer interaction: A review. *IEEE Trans. PAMI*, 19(5):677–695, 1997.

[17] Poesia Filter. http://www.poesia-filter.org/.

[18] T. Starner and A. Pentland. Visual recognition of American sign language using hidden Markov models. In *Proc. WAFGR*, pages 189–194, 1995.

[19] M. Su, H. Huang, et al. Application of neural networks in spatio-temporal hand gesture recognition. In *Proc. IEEE WCCI*, pages 2116–2121, 1998.

[20] P Viola and M Jones. Robust real-time object detection. *Int. Journal on Computer Vision*, 2002.

[21] T. F Wu, C. J Lin, and R. C Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–425, 2004.