HAND WASHING STEPS RECOGNITION SYSTEM

Ye Changhe, Zhang Haihan, Zhang Le

Institute of Systems Science, National University of Singapore, Singapore 119615

ABSTRACT

Keep hand-hygiene is one of the most important rules that health-care workers (HCWS) need to follow, especially in current situation, COVID-19 moves from epidemic to pandemic status all over the world, HCWS following the standard hand-wash steps strictly can prevent health care–associated infection (HCAI). In this paper, a real time video recognition application with Dense spacetime feature and Deeply learned feature combined is presented to detect people's hand-washing steps, which can guide HCWS to follow the most common hand-hygiene process - 7 step hand-washing process. 2 types of different recognition models (3DCNN and SVM) are used for feature extraction and action classification are tuned and stacked together to performance the hand washing steps recognition and the finial model achieved 84.1% accuracy.

Index Terms— Video Action Recognition, 3DCNN ,HOF, SVM

1. INTRODUCTION

Health care-associated infection (HCAI) is a major cause of death and disability worldwide Defective hand cleansing leads to poor hand decontamination. Obviously, when HCWs fail to clean their hands during the sequence of care of a single patient and/or between patients' contact, microbial transfer is likely to occur[1].

A research from showed that in developed countries, HCAI concerns 5%–15% of hospitalized patients and can affect 9–37% of those admitted to intensive care units (ICUs)[2]. However in developing country, because of under staffing, poor hygiene and sanitation, lack or shortage of basic equipment,the numerous viral and bacterial HCAI are transmitted and the burden due to such infections seems likely to be several times higher than what is observed in developed countries[1].

Thus, WHO released the WHO Guidelines on Hand Hygiene in Health Care at 2009. In this guide, 7 steps handwashing process have been seen as a standard hand-hygiene process, which contain under steps(See Figure 1), whole process need continue more than 30 seconds:

- **Step 1:** Palm to palm
- **Step 2:** Between fingers
- **Step 3:** Back of hands
- **Step 4:** Base of thumbs
- **Step 5:** Back of fingers
- **Step 6:** Fingernails
- **Step 7:** Wrists

Fig. 1. 7 steps of hand hygiene

However, HCWs to recommended hand hygiene procedures has been reported as variable, with mean baseline rates ranging from 5% to 89% and an overall average of 38.7%[1].

Clearly, an intelligent system which can recognize the hand-washing action may help HCWS adherence to the hand hygiene recommendations and reduce the HCAI.

2. BUSINESS UNDERSTANDING

Our objective in this system is to detect if human is following correct approach when they're washing hands, which means they're following a specific sequence of steps, and each step continues enough time. Base on the recognition result, system need to give user prompt like if they passed current step can continue next step or if they passed the whole hand-hygiene approach.

3. DATA UNDERSTANDING

In this project, we use self collect videos to apply the model. The video dataset contains total 659 short videos which can

be categorised to 7 steps.All videos were took in different angles and environments. The length of each video is about 2-5 seconds and each short video contains only one step(See Figure 2 & Figure 3)

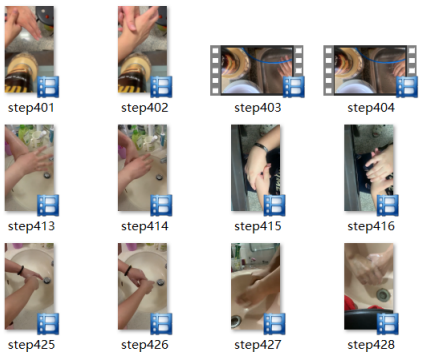

Fig. 2. video example of dataset

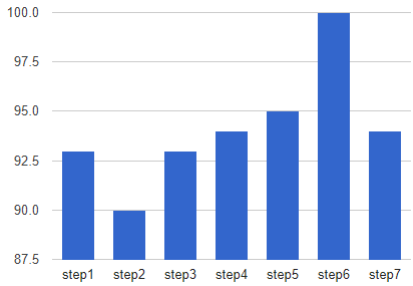

Fig. 3. dataset amount distribution

4. SYSTEM DESIGN

1. Summary

At design stage, we assume that in practice, one camera will be settled at the sink in the washroom, which can capture people's hand and the movement/action of hands. A GUI (See Figure 6) will be provided to assist with people to operate the application and also display the live video captured by camera. Once they start to wash their hands, at same time the video frames will be passed to system as input, and a pre-trained machine learning model will help to analysis the input frames therefore recognise the current hand actions. The GUI help to display the inference result and indicts people what should to do in next. the recognize result showing that a people washed his/her hands by a specified step sequence and the time of each step is long enough, then this whole hand washing process can be recorded as success. Here is a system architecture diagram in Figure 4.

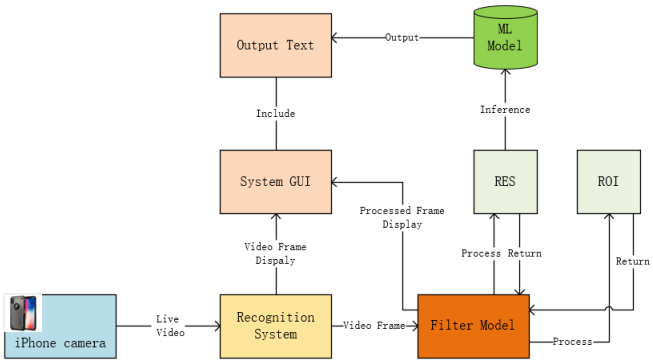

Fig. 4. System architecture diagram

2. Design Detail

First of all, system will get the real time video frames from camera, once the system gets 10 frames from video, it will resize frame into 32*32. Then this 10 frame short clip will apply to the action recognize model to predict if these 10 frames can be considered as step n (1 to 7).A threshold value is settled to prevent none required action be recognized as one of the required step. When the score of the prediction result is bigger than threshold value, this clip will be categorised to this action(step). Since we hope people could follow the correct steps and wash their hands with enough time, only if system successfully recognized step n for 5 times (totally about 4 seconds) people are allowed to go to next step. Figure 5 shows the whole process flow of the system.

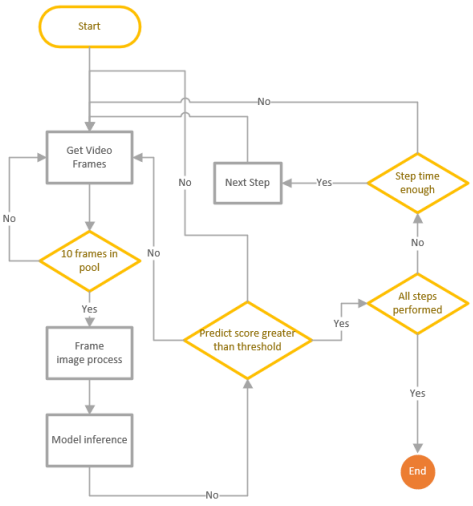

Fig. 5. Process flow diagram

Fig. 6. System GUI

3. Recognition Model Design

- Dense space time feature

  Firstly, we get a 10 frames clip from camera, then feature extraction is performed to extract useful information from each frame. We extracted the HOF (Histogram of Optical Flow) feature. Optical flow is the pattern of apparent motion of objects, surfaces, and edges in a visual scene caused by the relative motion between an observer and a scene.[4] Optical flow can also be defined as the distribution of apparent velocities of movement of brightness pattern in an image.[5] In HOF, angles between the optical flow vectors and the horizontal axis will be calculated, then project them into the corresponding histogram bin according to the angle, and adjust base on the amplitude of the optical flow (See Figure 7). Extracted features will be fed into SVM classifiers to perform classification.

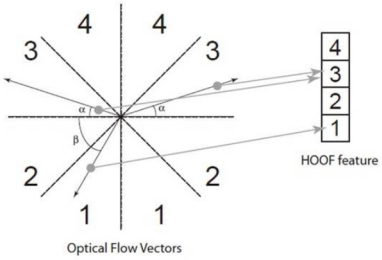

Fig. 7. HOF Calculation

- Deeply learned representation

  In this model, we feed video frames into 3D-CNN Model to do further feature extraction. 3D-CNN model extracts features from both spatial and temporal dimensions by performing 3D convolutions, thereby capturing the motion information encoded in multiple adjacent frames. The developed model generates multiple channels of information from the input frames, and the final feature representation is obtained by combining information from all channels(See Figure 8).Figure 3 showed the construction of our 3D-CNN model. As the business problem is a multi-label classification problem with 7 different component failures, the model will output the classification score directly.

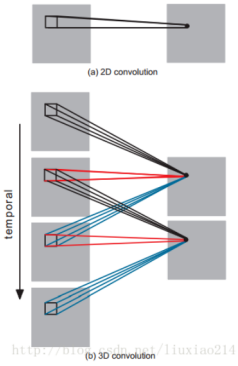

Fig. 8. Convolutions Process

```
def createBasicModel(input, nb_classes):
    inputs = Input(shape=input.shape[1:])

    x = Conv3D(32, (3, 3, 3), padding='same',
               kernel_initializer='he_normal',
               kernel_regularizer=l2(0.00001))(inputs)

    x = Activation('relu')(x)
    x = BatchNormalization()(x)
    x = Conv3D(32, (3, 3, 3), padding='same',
               kernel_initializer='he_normal',
               kernel_regularizer=l2(0.00001))(x)

    x = Activation('softmax')(x)
    x = MaxPooling3D(pool_size=(2, 2, 2), padding='same')(x)
    x = Dropout(0.25)(x)

    x = Conv3D(64, (3, 3, 3), padding='same',
               kernel_initializer='he_normal',
               kernel_regularizer=l2(0.00001))(x)

    x = Activation('relu')(x)
    x = BatchNormalization()(x)
    x = Conv3D(64, (3, 3, 3), padding='same',
               kernel_initializer='he_normal',
               kernel_regularizer=l2(0.00001))(x)

    x = Activation('softmax')(x)
    x = MaxPooling3D(pool_size=(2, 2, 2), padding='same')(x)
    x = Dropout(0.25)(x)

    x = Flatten(name='flatten_feature')(x)
    x = Dense(1024, activation='relu', kernel_initializer='he_normal')(x)
    x = Dropout(0.2)(x)
    outputs = Dense(nb_classes, activation='softmax')(x)

    m = Model(inputs=inputs, outputs=outputs)

    return m
```

Fig. 9. 3D-CNN model structure in this project

- HOF + 3D-CNN + SVM

  In this model, we combined the first 2 methods together. Feature extracted from 3D-CNN model will be used to train the SVM classifier, which output result of classification score, and at real time recognition stage, the HOF features will be feed into svm instead of raw data. please see in figure 10 for the summary of 3D-CNN model part.

4. System Integration

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 32, 32, 10, 1)	0
conv3d (Conv3D)	(None, 32, 32, 10, 32)	896
activation (Activation)	(None, 32, 32, 10, 32)	0
batch_normalization (BatchNo	(None, 32, 32, 10, 32)	128
conv3d_1 (Conv3D)	(None, 32, 32, 10, 32)	27680
activation_1 (Activation)	(None, 32, 32, 10, 32)	0
max_pooling3d (MaxPooling3D)	(None, 16, 16, 5, 32)	0
dropout (Dropout)	(None, 16, 16, 5, 32)	0
conv3d_2 (Conv3D)	(None, 16, 16, 5, 64)	55360
activation_2 (Activation)	(None, 16, 16, 5, 64)	0
batch_normalization_1 (Batch	(None, 16, 16, 5, 64)	256
conv3d_3 (Conv3D)	(None, 16, 16, 5, 64)	110656
activation_3 (Activation)	(None, 16, 16, 5, 64)	0
max_pooling3d_1 (MaxPooling3	(None, 8, 8, 3, 64)	0
dropout_1 (Dropout)	(None, 8, 8, 3, 64)	0
flatten_feature (Flatten)	(None, 12288)	0
dense (Dense)	(None, 1024)	12583936
dropout_2 (Dropout)	(None, 1024)	0
dense_1 (Dense)	(None, 7)	7175

Total params: 12,786,087
Trainable params: 12,785,895
Non-trainable params: 192

Fig. 10. Convolutions Process

Our system is mainly separated into four different modules:

- Video Capture Module: leverage iVcam app to allow the laptop connect to iPhone camera and use openCV library to do the video frames processing.
- Qt5 GUI: implemented by Qt5 to craft a user interface for user to operate the hand washing recognition application.
- Skin Mask: Existing module which can filter the background and only display the hand image.
- HOF Feature Extraction: Existing method which can extract hof feature from the video input.
- Inference ML Model: by tuning the existing c3d model and combine with the SVM to achieve best potential inference accuracy.
- Action Recognition and Process Module: developed by our own to full fill the whole logical flow process.

Module	Existing	Change
Video Capture	No	Integrated
Qt5 GUI	No	Developed
Skin Mask	Yes	Integrated
HOF Feature Extraction	Yes	Integrated
ML Model	Yes	fine-tuning
Action Recognition	No	Developed

Table 1. System Module List

5. EXPERIMENTAL RESULTS

For the model training part we mainly experimented three method and try to achieve the best performance of the predict accuracy:

- **Method 1:** HOF feature with SVM

  Here we extract the HOF feature from the video frames and use SVM to perform classification.

	precision	recall	f1-score	support
Step1	0.71	0.65	0.68	23
Step2	0.67	0.87	0.75	23
Step3	0.81	0.72	0.76	18
Step4	0.69	0.65	0.67	17
Step5	0.76	0.65	0.70	20
Step6	0.79	0.65	0.71	17
Step7	0.56	0.71	0.63	14
avg / total	0.72	0.70	0.70	132

Table 2. Accuracy score for SVM

- **Method 2:** Fine tuning 3D CNN

  Then we try with the 3D CNN to combine every 10 frames from each video into clips and then feed the clips to the machine learning model train the data for 100 epoch and from the fusion matrix we can see the accuracy improved a lot comparing with the pure SVM model.

Fig. 11. Training result for 3D-CNN

	precision	recall	f1-score	support
Step1	0.7000	0.9130	0.7925	23
Step2	0.8333	0.6522	0.7317	23
Step3	0.8824	0.8333	0.8571	18
Step4	0.8947	1.0000	0.9444	17
Step5	0.7059	0.6000	0.6486	20
Step6	0.9375	0.8824	0.9091	17
Step7	0.8000	0.8571	0.8276	14
avg / total	0.8153	0.8106	0.8072	132

Table 3. Accuracy score for 3D-CNN

- **Method 3:** 3D CNN + SVM

  In this end, we combined with two method together let the 3D-CNN model do the feature extraction and then stack the features with SVM classifier and we got finally 84.1% accuracy.

	precision	recall	f1-score	support
Step1	0.74	0.87	0.80	23
Step2	0.82	0.78	0.80	23
Step3	0.88	0.83	0.86	18
Step4	0.94	0.94	0.94	17
Step5	0.82	0.70	0.76	20
Step6	0.93	0.82	0.87	17
Step7	0.82	1.00	0.90	14
avg / total	0.85	0.84	0.84	132

Table 4. Accuracy score for 3D-CNN + SVM

after we get the model which has the best classification ability, we integrated it into our system there are some screen shot when we washing the hands with the camera on (see figure 11 to figure 14) and we can see the back-end console is printing out the inference result with the probability of each action recognized (see figure 15).

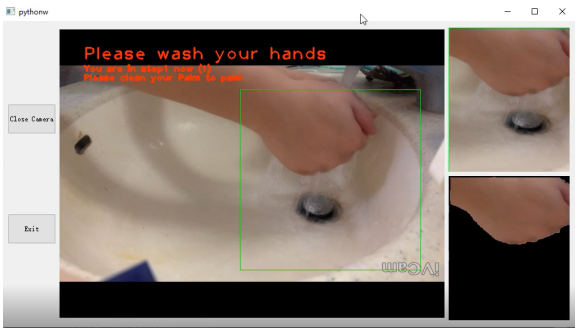

Fig. 12. System demo 1

Fig. 13. System demo 2

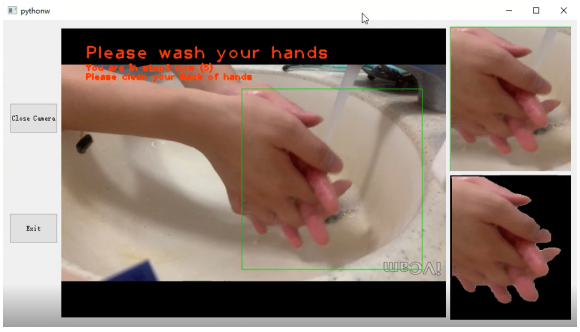

Fig. 14. System demo 3

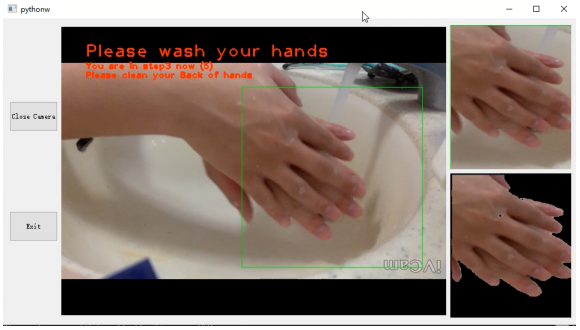

Fig. 15. System demo 4

```
Current stage: step3
this is None and the probability is 0.5067524137959061
[0.58487246 0.26987894 0.06417618 0.03836914 0.00645056 0.02590487
 0.01034785]
[0.58487246 0.26987894 0.26417618 0.03836914 0.00645056 0.02590487
 0.01034785]
Current stage: step3
this is None and the probability is 0.5848724553372179
[0.50859108 0.46188197 0.02327992 0.00111161 0.00285096 0.00130818
 0.00097628]
[0.50859108 0.46188197 0.22327992 0.00111161 0.00285096 0.00130818
 0.00097628]
Current stage: step3
this is None and the probability is 0.508591082863739
[0.62131374 0.3405636  0.01772274 0.0080659  0.00358287 0.00189524
 0.00685593]
[0.62131374 0.3405636  0.21772274 0.0080659  0.00358287 0.00189524
 0.00685593]
Current stage: step3
this is None and the probability is 0.6213137360608657
[0.2832601  0.11094683 0.00805329 0.03998397 0.04888777 0.48349476
 0.02537328]
[0.2832601  0.11094683 0.20805329 0.03998397 0.04888777 0.48349476
 0.02537328]
Current stage: step3
this is None and the probability is 0.4834947609024309
[0.57419119 0.20417058 0.01516693 0.07429449 0.0200681  0.06110811
 0.0510006 ]
[0.57419119 0.20417058 0.21516693 0.07429449 0.0200681  0.06110811
 0.0510006 ]
```

Fig. 16. Back-end output result

6. CONCLUSION

In this work, the real time human actions during handwashing process are recognised via a intelligence system, which can help guide HCWS follow the standard handhydrine process strictly during the work and reduce the chance of HCAI.

Two different methods of signal processing, namely the HOF feature extraction and 3D-CNN, are used for feature extraction, and 2 classifier: SVM and single 3D-CNN are used

to apply final classification. The model performances based on Table 2, Table 3 and Table 4.Results show that the single 3D-CNN approach outperforms the HOF feature + SVM classifier method by about 10% prediction accuracy, and the 3D-CNN feature + SVM classifier + HOF model approach out performs the single 3D-CNN model with about 4% prediction accuracy.

Best performance come from the third method is because that the combination of 2 features (3D-CNN features and HOF features), classifier can learn much better than only one type feature on video action classification. Currently the model is trained based on a very limited self-collecting dataset and the recognition of actions may still fail because of odd angles and other environment noise. Hence, the prediction can be further enhanced by extend the size of dataset in the future. Also, introducing more features like hand skeleton in training stage also can be expected to improve the prediction accuracy.

7. REFERENCES

[1] WHO Guidelines on Hand Hygiene in Health Care: Summary, 2009.

[2] World Alliance for Patient Safety. The Global Patient Safety Challenge 2005-2006 “Clean Care is Safer Care”. Geneva, World Health Organization, 2005.

[3] S. Ji, W. Xu, M. Yang, and K. Yu. 3D convolutional neural networks for human action recognition. PAMI, 2013.

[4] Burton, Andrew; Radford, John (1978). Thinking in Perspective: Critical Essays in the Study of Thought Processes. Routledge. ISBN 978-0-416-85840-2.

[5] Horn, Berthold K.P.; Schunck, Brian G. (August 1981). ”Determining optical flow” (PDF). Artificial Intelligence. 17 (1–3): 185–203. doi:10.1016/0004-3702(81)90024-2. hdl:1721.1/6337.