IMAGE PROCESSING

Independent Study Report

Video Analysis - Medical Sink Interaction

Advisor: Prof. Scott Acton

Name: Sharath Chand PV (ID: vp4pa)

Introduction

Videos are captured around the sink area in the selected care units of UVA hospital. The purpose of the study is to understand the characteristics of various types of interactions with the sink by analyzing and extracting insights from the videos. The information then can be used in understanding the spread of hospital specific infections. More specifically, it can also be used in understanding the correlation between the nature of interactions with the sink and the occurrence of infections.

Objectives

1. Develop a methodology to classify each incoming video based on the nature of activities in them
2. Extract information related to type of interaction with sink by analyzing the videos

Background

The cameras mounted above sink areas has motion sensing activation and are used to capture sink interaction videos. The cameras gets triggered 6 to 8 seconds before the appearance of any foreground object that gives us reference image for each video. The sample of videos used to develop initial methodology of analysis has 5 types of activities. First step in our analysis is to classify the videos based on the type of activities.

Following are the type of activities captured in the initial sample of videos.

Activity	**Count**
Hand washing	4
Medical Sink Interaction	5
Items Placed	5
Water Pouring	4
Water Running	3
Total	21

Since there is a possibility of more activities in the actual data, we have considered 4 classes i.e Handwashing, Medical sink interaction, Items placed and Others in our initial categorization methods.

Since the videos of all the rooms will be stored in a folder everyday through a batch job, we have developed a code (video_classification.ipynb) to extract and store the information of each video from the folder which can be later used to access the videos for the analysis. The code generates following tabular form and stores it in *video_data.csv* file. This also lets us analyze the frequency of sink usage in different time slots in different rooms.

Following table shows output of 10 files

Table 1.0

Day	Day_of_week	Month	RoomNo	Time_hh	Time_mm	Time_ss	Year
15	Fri	Jan	3132	10	24	13	2016
15	Fri	Jan	3132	14	44	18	2016
15	Fri	Jan	3132	15	47	38	2016
15	Fri	Jan	3132	19	15	7	2016
15	Fri	Jan	3132	19	15	44	2016
16	Sat	Jan	3132	9	59	18	2016
16	Sat	Jan	3132	11	32	30	2016
16	Sat	Jan	3132	11	33	30	2016
16	Sat	Jan	3132	13	5	10	2016
16	Sat	Jan	3132	17	38	29	2016

Methodology

The first step is to categorize the videos into different activity classes. We have developed two approaches. First approach is by identifying and analyzing foreground objects in the frames. Second approach is by generating features from the learned weights of fully trained convolutional neural network and using standard classifiers for classification jobs.

1. Foreground Object Detection

Since we can get our reference images due to motion sensing activation functionality of cameras and also there are not many moving objects in the video, we have decided to apply basic background subtraction to get foreground objects. Given the type of activities we have developed separate methods to identify each activity based on their characteristics. We have given detailed set of characteristics used to identify each of the four categories defined.

a. Items Placed on Sink body

The activity in these type of videos are either placing an item (plates, cups or glasses) on sink body or removing a previously placed item from the sink body. It is easy to identify these activities using differences between first and last frames.

b. Medical Sink Interactions

These videos have similar first and last frames. One of the distinct and useful characteristic in these activities is the gloves used for medical interactions. In almost all the medical interaction activities a purple colored glove is worn by the subject. That glove was used to categorize these set of videos

c. Handwashing

Similar is Medical Sink interaction type, these videos also have the similar first and last frames. By identifying foreground object which is not of purple color and has skin color can be categorized into washing set of activities

d. Others

Videos with activities such as Water running and pouring or with any other new type of activity shall be categorized into this class.

We have developed 3 different functions for identifying first 3 type of activities above, if a video gets negative flag from all the 3 functions then that will be categorized as “Other”. The functions that classify the videos into first 3 classes are:

1. Items Placed (function: get_first_last_diff())
    a. Get first and last frame from a video
    b. Get differenced frame from absolute difference of first and last frame
    c. Convert the differenced frame to grayscale then binarize at a threshold of 80
    d. Perform erode operation with a kernel of 5 x 5
    e. Determine if there is a significant difference between first and last frame from the pixel count
2. Medical Sink Interaction (function: glove_detector()):
    a. Identify frames with foreground using absolute difference with the reference frame
    b. If the frame has a foreground object, then convert it to HSV color scale
    c. Create a mask with range of purple color in HSV (color of gloves)
    d. Get key points that match with the purple mask using *bitwise_and()* function in openCV. A frame with Non-zero count of key points has a glove in it.

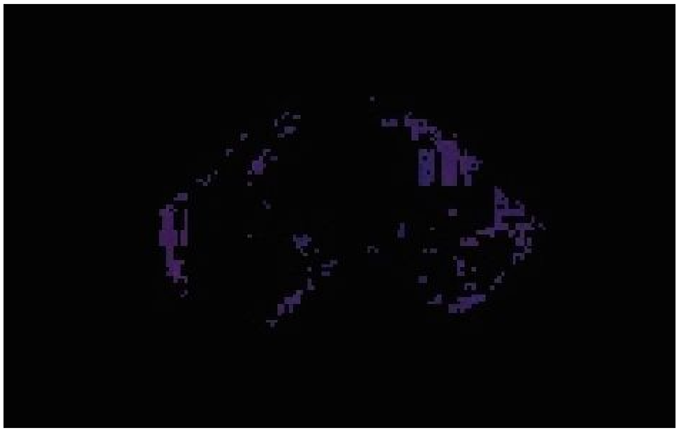


Fig 1: Detected glove

    e. Categorize the video based on number of frames with purple colored gloves in it

3. Handwashing (function: handwash_detector()):
    a. In this case, we considered first frame and last frame both as reference frames
    b. Identify frames with foreground using absolute difference with the reference frames. Apply morphological operations with a kernel of 5 x 5

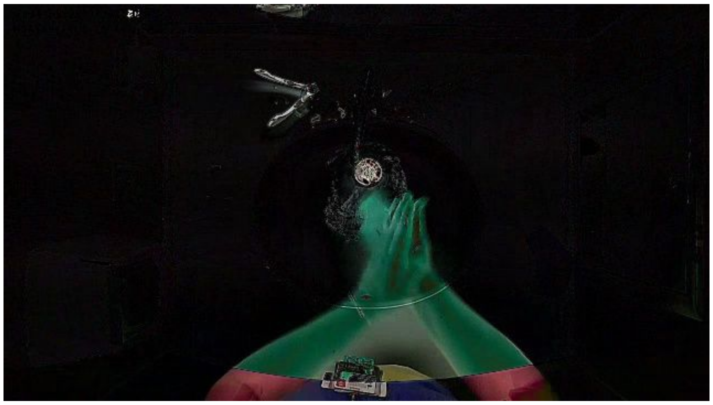


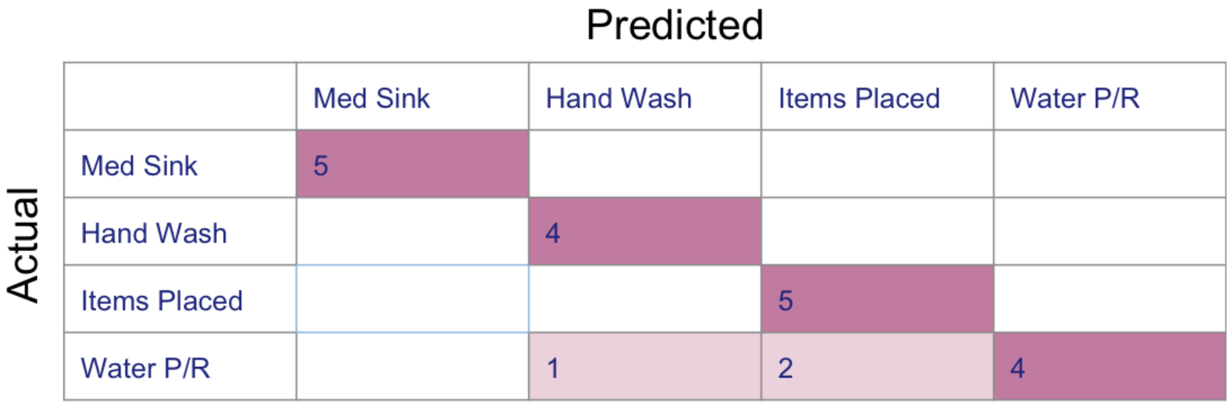
Fig 2: Foreground object - Handwash

    c. Considered as potential candidate frame for handwash if the frame has non-zero absolute difference with both the reference frames
    d. Categorize the video based on number of handwash frames
4. Others:
    a. Videos that don't fall into any of the above categories

There is a possibility that a video can have multiple activities in it. We have adopted soft classification in which a video can be categorized into more than one class. In the code, we used the csv file (Table 1.0) to read each video sequentially. The read video is sent through each of the classification function and given a positive flag if it satisfies the criteria of the function. It allows us to attribute more than one class to a video if it has multiple activities.

Results

We have tested our program on 21 sample videos. Following table summarizes our results



		Predicted			
		Med Sink	Hand Wash	Items Placed	Water P/R
Actual	Med Sink	5			
	Hand Wash		4		
	Items Placed			5	
	Water P/R		1	2	4

All the first 3 classes are being categorized correctly but 3 videos which are labelled as others are being classified as Handwash / Items Placed by our functions.

2. Convolutional Neural Nets

Convolutional networks (ConvNets) currently set the state of the art in visual recognition.

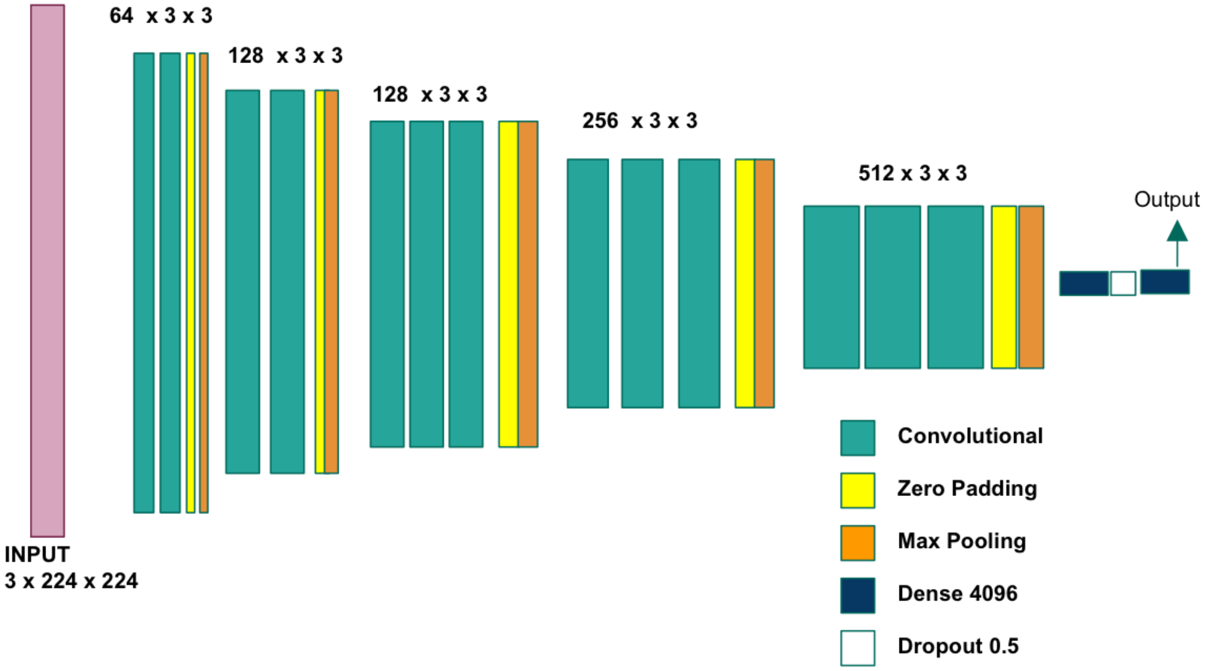
We have applied a novel approach in extracting features from videos using the weights of pre-trained convolutional neural networks. The approaches were discussed in detail in [1] and [2]. The extracted features were used to train standard classifiers such as SVMs or Random forests to classify the videos.

Transfer Learning

The weights of deeper hidden layers of pre-trained convnets are used to extract complex features from an input image. The image representations learned by networks that are trained on large scale datasets can be effectively transferred to other recognition tasks with smaller datasets. It has been successfully proved that the deeper layers has a ability to learn complex features and the kernel weights of them can be used on new input images to represent them in the form of those complex features.

We have used Visual Geometric Group's network popularly known as VGG16 network, one of the top performing teams in ILSVRC (Imagenet Large Scale Visual Recognition

Challenge), in our transfer learning task. The VGG16 network was trained 1.3 million images with 1000 output classes. Since our task is not the image classification, we have discarded the output layer and retained the network upto 2nd fully connected layer. The network used is shown below.



Frame Sampling:

We have developed a non-uniform sampling technique based on foreground object appearance. We have set no. of frames to be sampled “N” as a parameter. Our sampling method is as per below.

Frame 1 to foreground object appearance - 1 nos

Frames with foreground objects - N-2 nos

Last frame with foreground object to last frame of video - 1 nos

We selected N=15 after validating different values. There are two benefits of this sampling method. It avoids redundancy by not processing the frames with same information repeatedly and it helps speeding up the processing. A typical 50 seconds video has 500 frames (with fps 10 frames/sec) and processing all the frames can severely affect the processing speed.

Feature Extraction

The sampled frames of videos were fed through the above described Convnet. The features (a vector of 4096 length) were extracted from 2nd fully connected layer (FC). Before feeding the images to the network, they were mean subtracted and resized to 224 X 224 x 3.

Since 15 frames were sampled from each video, after the feature extraction from 2nd FC layer, each video can be represented as 15 x 4096 tensor. We have computed a mean feature of length 4096 for each video from the 15 image features. The mean feature was used as features in classification task.

Unlike learning tasks in convnets which requires a large dataset in our case thousands of videos, transfer learning method can be effectively applied with a couple of hundred videos. We have demonstrated that learning is happening on the features extracted by fitting a random forest classifier on the small set of videos we have.

The following table shows predicted probabilities and True class label for Handwash classifier.

Probabilities = [Pr (Video is not handwash), Pr (Video is handwash)].

Label = True class. We have considered One Vs Rest classification.

We can observe that the classifier has predicted higher probability of handwash for which the true class is handwash. But since we have tested the classifier on the same data which was used in training, we cannot expect similar performance on unseen data. But these predicted values shows that features are helpful in learning the differences between handwash and non-handwash videos. We have observed similar learning abilities in other classes of videos as well.

Handwash			Med Sink Interaction	
Probabilities	Label		Probabilities	Label
[0.93, 0.07],			[0.95, 0.05],	
[0.94, 0.06],			[0.97, 0.03],	
[0.34, 0.66],	Handwash		[0.97, 0.03],	
[0.88, 0.12],			[0.99, 0.01],	
[0.91, 0.09],			[0.97, 0.03],	
[0.93, 0.07],			[0.84, 0.16],	
[0.98, 0.02],			[0.97, 0.03],	
[0.95, 0.05],			[0.99, 0.01],	
[0.89, 0.11],			[0.94, 0.06],	
[0.86, 0.14],			[0.89, 0.11],	
[0.37, 0.63],	Handwash		[0.85, 0.15],	
[0.83, 0.17],			[0.85, 0.15],	
[0.9 , 0.1],			[0.21, 0.79],	Med Sink
[0.38, 0.62],	Handwash		[0.87, 0.13],	
[0.94, 0.06],			[0.24, 0.76],	Med Sink
[0.95, 0.05],			[0.16, 0.84],	Med Sink

[0.24, 0.76],	Handwash		[0.86, 0.14],	
[0.88, 0.12],			[0.13, 0.87],	Med Sink
[0.86, 0.14],			[0.18, 0.82],	Med Sink

Handwash details

We also have developed a function that takes videos that classified as Handwash as input and generates details such as the duration of handwashing activity, start and end frame etc. The duration was estimated from foreground object appearance and the fps (frames per sec) of the video. The same method can be applied to other categories to estimate the duration of activity.

Sample Output for Handwash videos:

```
handwash start frame is 76 and end frame is 572
Total length of handwash is 49.000000 seconds for 3132 - Fri Jan 15 15-47-38 2016.mp4
handwash start frame is 29 and end frame is 212
Total length of handwash is 18.000000 seconds for 3135 - Fri Jan 15 11-23-35 2016.mp4
handwash start frame is 70 and end frame is 250
Total length of handwash is 18.000000 seconds for 3135 - Fri Jan 15 18-43-35 2016.mp4
handwash start frame is 109 and end frame is 335
Total length of handwash is 22.000000 seconds for 3135 - Sat Jan 16 09-41-55 2016.mp4
```

Challenges

The transfer learning that we experimented here requires gpu computation functionality to implement in practice. It took approximately an hour for 21 videos each with 15 sampled frames, but with gpu (5gb, Nvidia Tesla k20c) it can achieved within 5 minutes.

The videos used to develop the programs are from 2 rooms, with same mounting range and angle and belong to only 5 different activities. In practice there could be many variations, mounting angle, range, background lighting, more number of activities (approximately 20 different classes). In that setting it would be difficult to develop functions for each class.

Conclusion

We believe that further exploring the transfer learning technique could yield better results in the setting where many variations as described above are expected. But transfer learning technique helps only in classifying the videos into their labelled classes. It still requires hand labelling of at least 500 videos or 50 videos per class to allow our classifiers (SVM or Random forest) to learn better. Once the classification is done we have to apply Image processing techniques to understand and analyze the activities further.

References

[1] D.C.Ciresan, U.Meie and J.Schmidhuber. Transfer learning for latin and chinese characters with deep neural networks. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–6. IEEE, 2012.

[2] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolu tional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1717–1724, 2014.

[3] VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION. Karen Simonyan∗ & Andrew Zisserman+Visual Geometry Group, Department of Engineering Science, University of Oxford. arXiv:1409.1556 [cs.CV]

Code

We have used openCV (cv2 package in python), skvideo and numpy packages in python to implement our first approach which is detecting foreground objects. For transfer learning, we have used deep learning packages Theano, Keras and standard classifiers from Sci-kit learn.

video_classification.ipynb - This code prepares a csv file, calls functions on each video to categorize them into one or the more classes and writes the output to the same csv.

transfer_learning_videos.ipynb: This code reads videos from the files aggregated in the csv, frame sampling, creates features from the output of fc2 layer of vgg16 network and fits standard classifier (random forest)

handwash_details.ipynb: This code reads Handwash videos and calculates the duration of washing activities, start and end frames of the activity.