Institute of Systems Science
National University of Singapore

MASTER OF TECHNOLOGY IN INTELLIGENT SYSTEMS

Graduate Certificate Online Examination

Subject: ***Intelligent Sensing Systems***

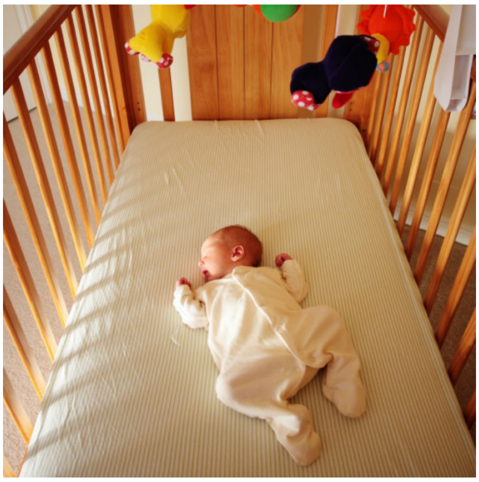
Sample Examination Questions

Graduate Certificate Intelligent Sensing Systems (ITSS) Sample Paper

Question 1

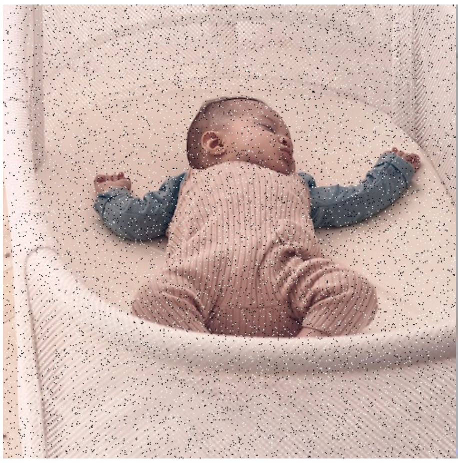
You are working in a MNC that is considering to build a smart baby monitoring system, which monitors baby's behavior at crib in order to avoid any unfortunate incident. However, it is not always possible for parents to monitor how their babies sleep. And therefore, your company aims to build a system that can monitor baby sleeping without any manual involvement. The system should continuously report the status of the baby at crib. Your company has a device ready for this purpose. The device comes with a HD colour night vision camera captured at 24 frames per second (RGB image), a microphone and a speaker. It also has a gooseneck clamp, as illustrated in the following figure.

You are tasked to build an AI that can discriminate the below 6 scenarios: no baby, baby asleep, baby awake, baby crying, face covered and baby on stomach. Your AI should work throughout the days and under diverse type of room setup. The figure below shows the image output from the device.

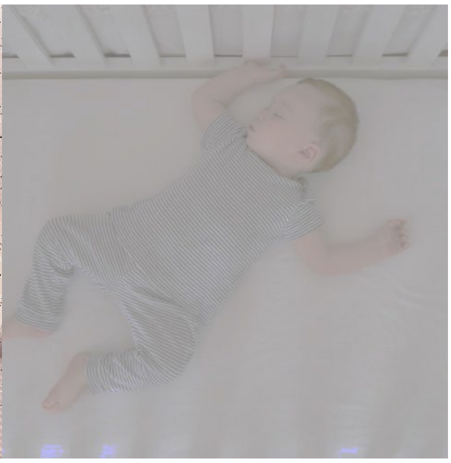


a. Before you start to develop any machine learning model, you are requested to perform segmentation on baby without spending any time on creating binary mask. Ideally, the boundary delineated on baby should use as fewer points as possible. (i) Propose a solution to achieve the above and (ii) specify the suitable image input (size, colour space, value range and value type) for your solution? Explain why do you use the specific algorithm(s) in your solution, and why the specified image input is the better option.

b. Based on the dataset you have (images of baby at crib, some images with no baby at crib), your colleague has built a deep learning model to classify a three-class scenario: lying baby, sitting baby, no baby. He has found that his model often fails to classify correctly in images like the below, and he is seeking your help. For each situation, you are requested to detail the issue, propose and explain your solution to rectify the image. Furthermore, you are asked to propose solution (with steps specified) that can make his deep learning model work better against each situation without modifying the deep learning model.



Situation 1



Situation 2

c. You just had a lengthy discussion with your senior data scientist. Both of you have agreed to build a deep learning model that accepts 48 frames of images (2-second video) as input to classify the required 6 scenarios. However, both of you have yet to reach a conclusion on how every single frame should be processed before it is consolidated as the input to deep learning model. Should the frames be fed in its entirety to the net (without any pre-processing), or should we run a baby detection on each frame, crop out and resize the region of interest and feed the 48 processed images to the net? Detail the advantages and the disadvantages of each approach and state your preference. Explain your decision.

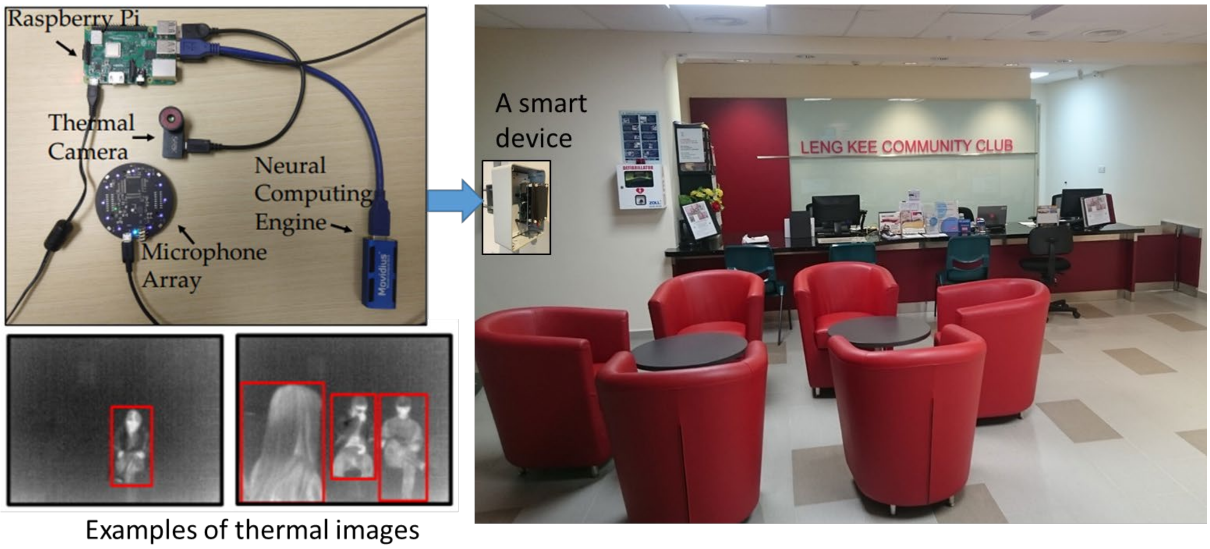
d. After building the model to distinguish the required six scenarios, you found that the trained model is not so capable in differentiating between "baby awake" and "baby crying". Given the provided device, what would you do to improve the accuracy? What would be the change in the input to deep learning model?

Question 2

Community clubs are common spaces in Singapore for people of all races to come together, build friendships and promote social bonding. Intelligent sensing technology can help community clubs not only just for ensuring public safety in the public place, but also for the purposes of promoting healthy life. To reduce the risk of the recent COVID-19 transmission of among residents, a few precaution measures have been implemented by this community club for its operations. One of the precaution measures implemented by this community club is to avoid overcrowding in the reception waiting area. In addition, the coughs sound passively captured in the public place along with the human counts together contain important information about the intensity of respiratory risks.

You are engaged by a community club to develop various solutions to address its needs. Answer the following questions.

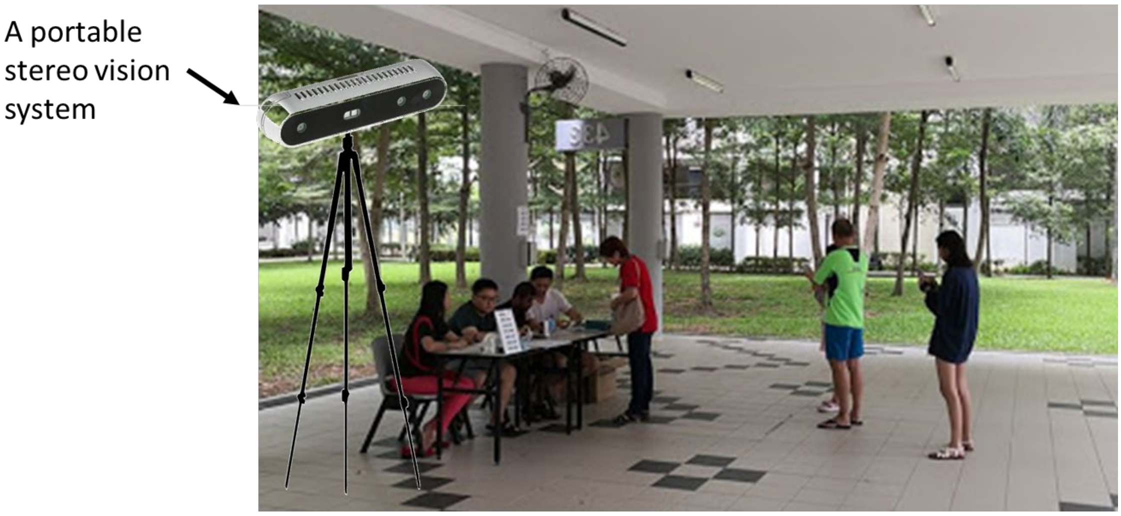
a. You are asked to develop an intelligent sensing system to provide following two objectives. (1) Provide a real-time **human counting system** that reminds residents not to enter the reception waiting area if it is over-crowded (e.g., 6 persons). (2) Provide a **real-time** "**cough index**", which is defined as the number of coughs per person per minute. The snapshot of the community centre reception waiting area and a set of equipment is provided for you as illustrated in Figure 1. You also can recommend additional equipment (if necessary). To provide privacy protection, no RGB camera is provided. In your proposed solution, you need to (1) **describe the system architecture** of your proposed solution, such as data acquisition and analytics; (2) **describe what vision/audio analytics methods** need to be applied to achieve the aforementioned two objectives.



Examples of thermal images

Figure 1. A snapshot of community club reception waiting area (right photo), examples of thermal images with detected human objects (bottom left photo), and the set of equipment provided for you (top left photo), which contains Raspberry Pi that is connected to thermal camera, microphone array, neural computing engine.

b. Every Singapore household can receive a free 500ml of zero-alcohol hand sanitiser. A temporary collection counter is set up for hand sanitiser collection in the community club. To avoid overcrowding, you are asked to develop a portable stereo vision system to measure the distance between queuing people to ensure there is enough space in between people lining up in queues. Your vision system consists of two calibrated cameras; they have the same focal length 20cm, a baseline 10cm, the pixel size in the camera is assumed to be 0.1cm/pixel; each image has a resolution of 1920*1080 pixels. **Evaluate what would be the maximum distance it can measure.** In order to monitor a bigger space and a larger distance, your camera vendor recommends you to purchase another product with the same stereo camera models but the two cameras are configured with a larger baseline distance (e.g., 15cm) between them. Do you agree with this camera vendor's suggestion? Explain your answers.



A portable stereo vision system

Figure 2. A snapshot of temporary hand sanitiser collection counter with a few people in a queue. A portable stereo vision system is deployed to ensure the safe distancing.

c. A table cleaning robot has been deployed by the community club to clean office facilities four times daily, as illustrated in Figure 3. You are asked to develop an intelligent vision system for this robot to recognize the surface materials and the table litters so that the cleaning robot can adjust its cleaning tool and path planning strategy. **Propose the image feature extraction method** to recognize different table surface materials (illustrated in Figure 4) and different litters on the table (as illustrated in Figure 5), respectively. In your answers, you need clearly describe (1) feature extraction and representation, and (2) feature modelling.

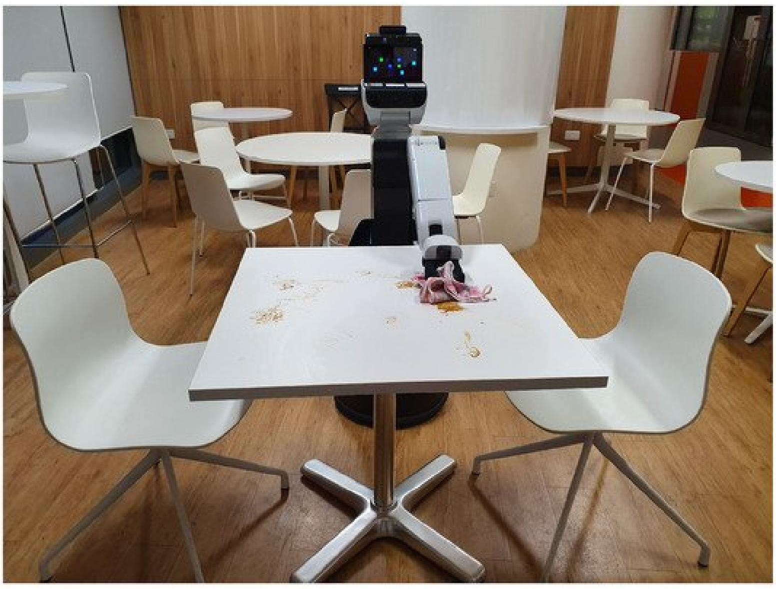


Figure 3. A table cleaning robot.

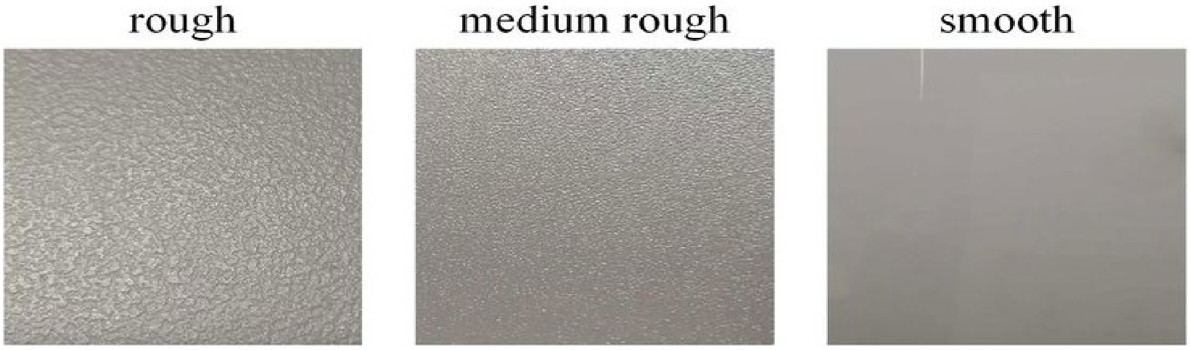


Figure 4. Examples of different surface materials.

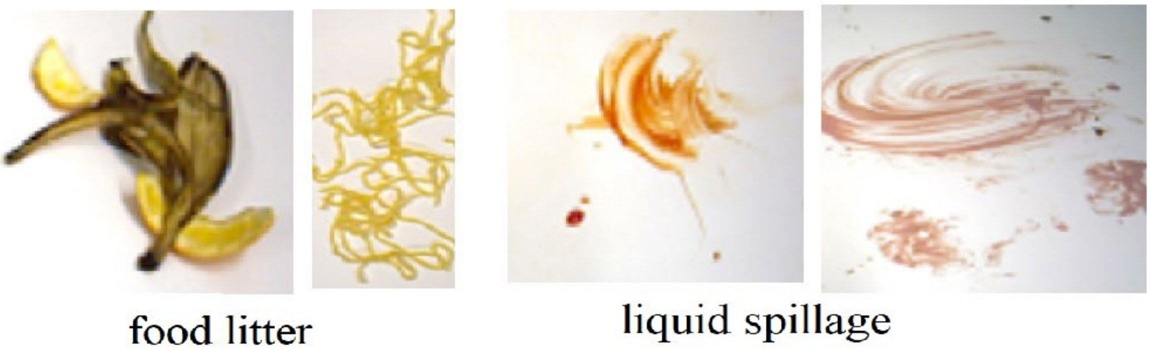


Figure 5. Examples of solid food litter and liquid spillage.

d. Residents have various types of exercises in the sports hall of the community club. To provide a better exercise facility resource planning, you are asked to develop a human action recognition system for four categories (i.e., basketball, running, badminton, walking) using the CCTV video footage. The first model A is a 3DCNN model, which uses the raw frames directly as the input to the machine learning model. The second model B is a Motion-2DCNN model, which applies the optical flow method on the input video sequence to calculate motion map for each two consecutive frames, then uses motion maps as the input to the machine learning model. Details of these two models are presented in Table 1. The machine learning model complexity is a key concern for real-time video analytics. **Evaluate these two models to identify which model is more**

complex, by comparing their output shapes, the numbers of trainable model parameters. Fill in your answers in Table 2. Furthermore, for the model with more trainable model parameters, **recommend one video pre-processing method** to further reduce its number of trainable model parameters, without modifying the model architectures provided in Table 1.

Table 1. Two machine learning models for human action recognition.

	Model A (3DCNN)	Model B (Motion-2DCNN)
Input video clip	A short video clip with 10 grey-scale frames, each frame has a resolution of 320*240 pixels.	
Convolution layer	A **3D convolution** layer with 16 filters, each has a kernel size of (3, 3, 3), 'ReLu' activation function, padding 'same'.	A **2D convolution** layer with 16 filters, each has a kernel size of (3, 3), 'ReLu' activation function, padding 'same'.
Flatten layer	A flatten layer	
Dense layer	A dense output layer with 4 nodes, 'Softmax' activation function; to recognize 4 categories of actions.	

Table 2. Fill your answers in the unshaded cells.

Layer Type	Activation	Kernel Size	Output Shape	Number of model parameters
Model A				
Input frames	-	-	(320,240,10)	-
3D convolution layer	ReLU	(3,3,3)		
Flatten	-	-		-
Dense	Softmax	-	4	
Model B				
Input motion maps	-	-	(320,240,9)	-
2D convolution layer	ReLU	(3,3)		
Flatten	-	-		-
Dense	Softmax	-	4	