|  |  |  |  |
| --- | --- | --- | --- |
| Question | Q1 | Q2 | Q3 |
| Marks | 6 | 8 | 6 |
| Total | 20 | | |

**Full Name (as in course registration): KAR CHAUDHURI ANIRBAN**

**Date: 11/04/2021**

**Course:** **Real Time Audio-Visual Sensing and Sense Making**

**Question 1.** For each question, select the **single** most appropriate answer, and **provide your justification** (no longer than two sentences).

a). Motion estimation method is used to estimate motion feature between two consecutive frames. Which of the following operation would **INCREASE** the dimension of the obtained motion features?

1. Apply the sub-pixel estimation, not the integer-pixel estimation, which is used in the block matching method. **(increases)**
2. Reduce the block size, which is used in the block matching method. **(increases)**
3. Increase the number of levels of the pyramid, which is used in the optical flow method. **(increases)**
4. All of above.

**Answer:** D.

**Justification:** Sub-pixel estimation results in a pixel being sub-divided into further pixels, reducing individual block size results in greater number of blocks for a frame while increasing number of pyramid levels increases loops of computational layers of Lucas-Kanade method. All three methods increase dimensions of obtained motion features.

b). Which of the following proposals is **feasible** for recognizing a single person dancing in an indoor dancing studio?

1. Generate a motion history image from the dancing sequence, followed by an image classification approach. **(feasible)**
2. Estimate motion field for every two consecutive frames, followed by a sequence classification approach. **(feasible)**
3. Extract human body shape from each individual image, followed by a sequence classification approach.**(not feasible)**
4. All of above.

**Answer:** A.

**Justification:** Extracting human body shape does not tell us displacement of human body, motion field of two consecutive frames for sequence classification is time consuming and sometimes may not be accurate if a person is not dancing at intervals. Motion history image that looks changing shape of object where pixel intensity is function of motion history will help us classify single label for an image easily and most feasible.

**Question 2.** In public venues, crowd size is a key indicator of crowd safety and stability. Monitoring the people number and crowd density levels are important. You are engaged by a security surveillance company to develop a security solution for New Year countdown party at the Sentosa beach, as illustrated in Figure below. The proposed system is required to have the following two functionalities: (a) human counting (crowd density); (b) stage event classification (sing, dance, etc). The camera is mounted at the entrance facing the stage (a snapshot of the CCTV image is illustrated in the right image). You can draw a flow chart to justify your proposed system.



**Crowd Counting (Human face tracking) - Tracking using Siam FC, a deep learning tracking method that can be used for multiple objects (human faces):**

1. A target in previous frame & a search region to be defined
2. Output: A score map, the position of maximum score relative to center of score map, multiplied with stride of network, gives displacement of target from frame to frame
3. Convolutional layer feature extraction maps involved for human face detection
4. Ground truth score mapping: 1 at pixel locations where human faces within certain radius and 0 elsewhere
5. Metrics for multiple object tracking: object positions to be detected using bounding boxes, intersection to be detected and determined using IOU calculation and if there is no intersection, Euclidean distances to be calculated between humans detected to determine density

**Stage Event Classification:**

1. Zoom into stage area, estimate camera motion to remove camera optical flow and try to stabilize camera where possible
2. Motion history image: pixel intensity is a function of motion history at stage performers location, where brighter values correspond to recent motion and used to extract motion field vectors for classifying human activity at stage
3. **Classifier: 3D Convnet, use early fusion**

**Question 3.** You are engaged by an electronic entertainment company to develop a human tracking system. Given the search reference image in Table 1, and the object template illustrated in Table 2, apply the *mean square error* (MSE) method to determine the best matched image region (integer pixel accuracy) in the search reference image. Show your calculations to justify your answer.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 1 | 2 | 2 |  | 1 | 2 |
| 3 | 2 | 3 |  | 2 | 3 |
| 4 | 4 | 4 |  | Table 2. The object template. The numbers are gray-scale intensity values. | |
| Table 1. The search reference image. The numbers are gray-scale intensity values. | | |  |
|  | | |  |  | |

Stride length one of in 3x3 image, **subtracting OOI pixels strength and then squaring the values.**

Step 1:

|  |  |
| --- | --- |
| 0 | 0 |
| 1 | 1 |

Step 2:

|  |  |
| --- | --- |
| 0 | 0 |
| 1 | 1 |

Step 3:

|  |  |
| --- | --- |
| 1 | 0 |
| 4 | 1 |

Step 4:

|  |  |
| --- | --- |
| 1 | 1 |
| 4 | 1 |

**MSE calculation:** Minimum error for stride 1 and 2. Thus, the best matched candidate region is given below.

|  |  |  |
| --- | --- | --- |
| 1 | 2 | 2 |
| 3 | 2 | 3 |