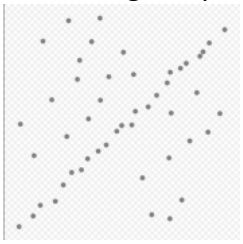
# Computer Vision 22-23
# Sheet 2

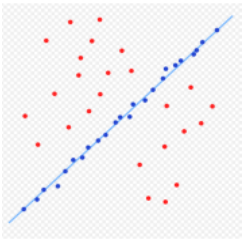
## Features

**Q1**) Using RANSAC for fitting a translational model, suppose we know that 30% of our data is outliers. How many times do we need to sample to assure with probability 80% that we have at least one sample being all inliers?

**Q2**) Given the following 2-D point cloud, it is supposed to fit an optimal line model to it.



(a) Original point cloud



(b) Line model after fitting

Describe how you will use **RANSAC** algorithm for the aforementioned task.

**Q3)** ***RANSAC*** (RANdom SAmple Consensus) is an algorithm for fitting a model to very noisy data. Its basic idea is to fit models to small samples of data, and then look for "consensus" from the entire dataset. Complete the following pseudo-code for RANSAC

**INPUT:**
data - a set of observations
model - a model that can be fitted to data
s - the minimum number of data required to fit the model (fixed for a given model)
t - a threshold value for determining when a data point fits a model
d - the number of "consensus" data points required for a "good" model

**OUTPUT**: BestModel
FOR (a number of iterations)
    CurrentSet := sample a data subset of size s
    CurrentModel := model fitted to CurrentSet
    ConsensusSet := CurrentSet
    FOR every data point not in CurrentSet
        ……
    END FOR

```
        IF #(ConsensusSet) > d
        ……..
        END IF
    END FOR
    RETURN BestModel
```

**Q4)** Consider 12 points are painted on a dark plane. 5 points are colored pure red, 5 are colored pure green, and 2 are colored white. Two cameras (stationed with a translational shift) take pictures of the plane. One camera is equipped with a red filter (doesn't see green color) and one camera is equipped with a green filter (doesn't see red color).

- **i.** ***Which*** points does each camera see?
- **ii.** Do you think that enough point correspondences are available to recover a transformation (homography) between the two images? ***Explain*** why.
- **iii.** ***Explain*** an algorithm that can compute the set of good point correspondences between the two images. Define the used parameters of the algorithm, and how the value of each parameter should be selected.

**Q5)** Suppose you have a large collection of photos from your trips, including photos of yourself with other people, photos of other people without you, as well as photos without people. Suppose you also have a template of your face, which consists of a small image of your frontal face. ***Describe*** a fully automatic algorithm that uses what you have learned in class to solve this "template matching" problem by finding the subset of the images that contains un-occluded frontal faces of you irrespective of their location, orientation and scale. (Hint: you don't have many example images of yourself except the template image).

## Object Recognition

**Q1**) Explain a scheme for accomplishing skin color recognition in images with the following emphasis:

- **i.** Explain what kind of low-level feature extraction you can use.
- **ii.** At the level of classifier, explain how the system can use Bayesian Framework to incorporate expert knowledge such as the skin color frequencies in images.

**Q2**) You are working for a special effects company and want to have real actors inserted into a computer-generated world. One way to do this is to film the actors in front of a blue screen, so that the background can easily be removed and substituted with a computer generated one. Due to lighting variations, the background will not appear pure blue, so checking for a single intensity value might not work well. Instead, you decide to model the distribution of background colors as a probability function. You are given training images T with pixels labeled as foreground (F) and background (B).

Describe a model based on Bayes Rule that given a new image can be used to recognize any pixel to be either background or foreground.

**Q3**) Describe the Bag-Of-visual-Words (BOW) image representation technique, stating what is meant by a visual word. Show how the BOW can be used for the problem of object recognition.

**Q4) Sketch** a general architecture of the Siamese Network with triplet loss used for face verification. Briefly **explain** the training procedure in terms of input, the learning objective and the loss function used.

**Q5**) In a Content Based Image Retrieval (CBIR) system, the purpose is to retrieve images from a large database using query by image content.

- **i.** Define the notion of the "semantic gap" in the context of systems for content based image retrieval.
- **ii.** Consider training two different ConvNets on a dataset with 10 different classes. After querying with the same cat image, the ranks of the retrieved images results as follows for the two models respectively:
  Model 1: (cat , dog, dog, cat, cat, car,….)
  Model 2: (cat, cat, cat, dog, car, car,….)
  If the database contains only 3 cat images, use the Average Precision (AP) metric to evaluate both queries and determine which model is more accurate.

**Q6)** Suppose you would like to build a car detection model to use at your garage. You have a database D = $\{x_1,...,x_N\}$ of the N cars allowed to enter your garage. A training example for your network is a triplet of images ($x_1$, $x_2$, $x_3$) where $x_1$ and $x_2$ are different images of the same car, and $x_1$ and $x_3$ are images of different cars. The network is supposed to verify any car by computing encodings $f_i$ and $f_j$ to any given two images $x_i$ and $x_j$ where pictures of the same car should have similar encodings.

- **i.** ***Sketch*** a general architecture of this network showing how it would determine whether a car in front of your garage should be allowed to enter.
- **ii.** Briefly ***explain*** the training procedure in terms of input, the learning objective and the loss function used.

## Object Detection

**Q1**) We want to build a safety system for cars that detects pedestrians from a single camera image. We have a large dataset of images some with pedestrians labelled by a bounding box. We want to be able to detect and locate pedestrians regardless of their distance and apparent size in the image.

**i.** Compare between R-CNN, Fast R-CNN and Faster R-CNN object detection and localization algorithms in terms of (use tabular format):

**1.** Used method of region proposals.

**2.** Number of ConvNets used for feature extraction.

**3.** Object classification function.

**4.** Speed.

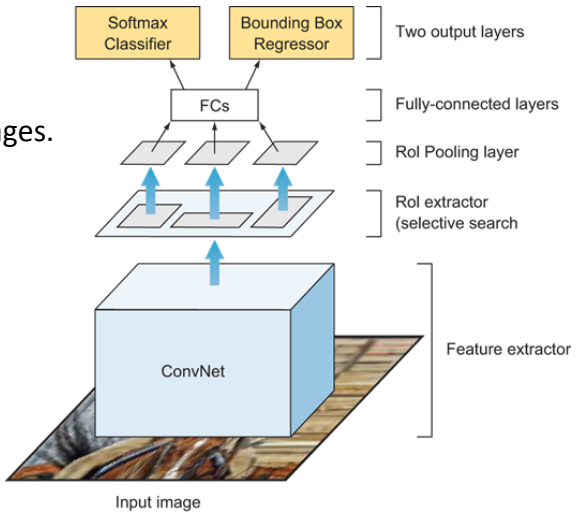
**iii.** Explain (in 4 statements) the YOLO algorithm used for real-time object detection and localization indicating how the system deals with size variations and the possible presence of multiple pedestrians.

**Q2**) The given diagram shows the Fast R-CNN network for object detection and localization in images.

**i.** Explain the function of the ROI pooling layer.

**ii.** Sketch a modified diagram to represent the Faster R-CNN network.

**iii.** Explain (in points) the main changes between the two models.

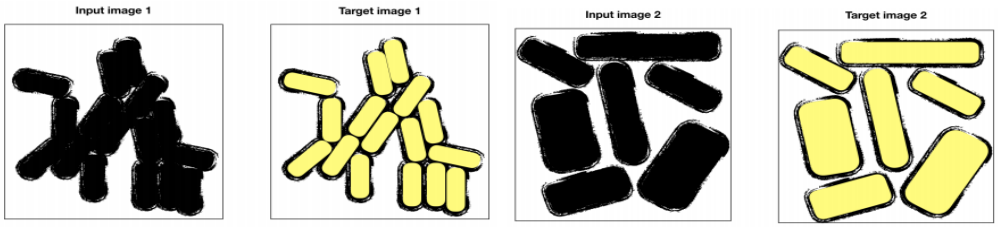



**Q3) Answer Briefly the following questions:**

1. Among the family of region-based object detectors, ***what*** makes the R-CNN very slow for the task of object detections?
2. Mention **One** example of a region-based CNN model that succeeded to enhance the speed of object detections of R-CNN. ***Explain*** how.
3. ***Explain*** the function of Region Proposal Network (RPN) as part of the Faster R-CNN object detector.
4. ***What*** is the typical output size of YOLO as a single-shot detector? ***Show*** what each part of the output indicates.
5. ***Explain*** a modified format of the YOLO output size that allows it to detect multiple classes in each grid cell.

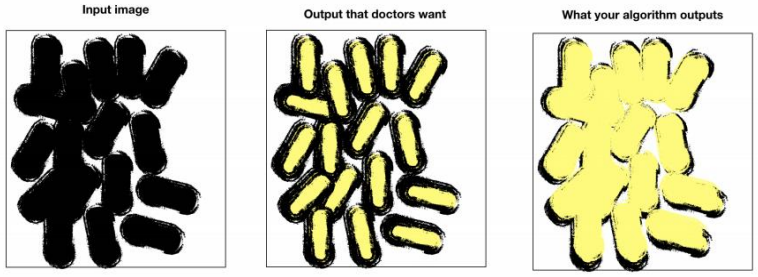
## Image Segmentation

**Q1**) You have been hired by a group of health-care researchers to solve one of their major challenges dealing with cell images: determining which parts of a microscope image corresponds to which individual cells. Given examples of input images and the corresponding target images as follows, the input images are taken from a microscope, and the target images have been created by the doctors. Your task is to implement an automatic model to achieve this task, where each pixel of the target image is labeled as 0 (this pixel is not part of a cell) or 1 (this pixel is part of a cell).




- **i.** Name the title of this computer vision task.
- **ii.** Design a network based on CNN to implement it (Sketch a general architecture diagram in terms of input, core and output components).

After finishing your model, the doctors are very unhappy, and argue that they would like to visually distinguish cells. They show you the following prediction output, which does not clearly separate distinct cells. Describe how you would correct your model to satisfy the doctors request.




- **iii.** Mention the title of this modified computer vision task.
- **iv.** Suggest the needed modification in the target images provided by the doctors in order to support the training of your model.

**Q2) Mean Shift** is a popular method used in computer vision for problems such as image segmentation and object tracking.

- **i.** **Describe** the main steps of how Mean Shift is used to perform image segmentation, including the parameter(s) that must be set in order to use Mean Shift.

ii. **Comment** on the sensitivity of the algorithm to its parameter value(s); i.e., what are the effects of setting the parameter(s) to values that are too big or too small.

iii. **Explain** the advantages of using the Mean Shift algorithm over K-means clustering for image segmentation.

**Q3)** Assume that you have been given images of the road taken from a camera mounted on the front of a car. Your task is to detect and segment cars using instance segmentation. Design a network based on CNN to implement this task. ***Sketch*** and ***explain*** a general architecture diagram in terms of input, core and output components.