CS376: Computer Vision: Sample FinalTerm

1. The exam is 90 minutes.
2. You are allowed one page of hand written notes. No calculators, cell phones, or any kindof internet connections are allowed.
3. The sample final-term is not representative of the true length or the point break-down of the final final-term. It is intended to provide you an idea of the range of topics and the format of the mid-term.

1 Multiple Choice

Each question has ONLY ONE CORRECT OPTION and is worth 2 points. Please draw a circle around the option to indicate your answer.

1. What are direct applications of generative models.

- (a) Synthesizing new images.
- (b) Generating training data for supervising neural networks.
- (c) Serve sub-modules for other tasks.
- (d) Image classification.

Solution: (a,b,c)

2.What of the following is irrelevant to the task of semantic segmentation?

- (a) TextonBoost.
- (b) Conv-Deconv.
- (c) Dilated-Conv.
- (d) Object Proposal.

Solution: (d).

3. What of the following influence the performance of k-nearest neighbor classifier? (Multiple choice)

- (a) the size of k.
- (b) the distance metric used to compute the nearest neighbor.
- (c) the size of the training dataset.
- (d) the size of the testing dataset.

Solution: (a,b,c)

4. What of the following is not true for boosting?

- (a) It combines weak classifiers.
- (b) It allows early-stop for efficient evaluation.

- (c) It is a parametric technique.
- (d) Its formulation is identical to that of a support vector machine.

Solution: (d)

5. What task can be solved if you have a perfect instance segmentation approach?

- (a) object detection
- (b) semantic segmentation
- (c) image classification
- (d) human pose estimation

Solution: (d)

6. Given a color image of resolution $227 \times 227 \times 3$, suppose you apply a linear filter of resolution 11×11 with stride 4 and no padding, then what is the resolution of the output image?

- (a) $55 \times 55 \times 3$
- (b) $33 \times 33 \times 3$
- (c) $54 \times 54 \times 3$
- (d) $56 \times 56 \times 3$

Solution: (a)

2 Long Answer

11.(9 points) Recognition
1. Suppose you want to classify textures in images. You have L images of texture with annotation to K classes, and you would like to use the Bag of Visual Words[1] for feature extraction. Describe the stages of such an algorithm, elaborate on the entire train and test stages.
Solution: The first stage will be to learn a dictionary of visual words.
Train:

- Extract M features from the images (note that M should be quite large) at random. These features could be SIFT descriptors, texture patches, Gaborresponse, laws filters or any other features that may capture the variance of texture patches.2.Learn a dictionary of words using unsupervised learning algorithm such as K-Means (or any more advanced methods such as sparse dictionary learning). Note that you need to define the dictionary size before this stage.
- The second stage would be to learn a specific category representation using a histogram of visual words for each category and afterwards a classifier of some sort (like SVM or KNN).

Test: For each test image, extract features, assign each feature to the nearest word in the dictionary using NN classifier, build the histogram of visual words and classify using the classifier from the train stage.
2. Suppose you want to discriminate between images of cats vs. images of cars. Suggest an algorithm based on dimensionality reduction that does not use local feature extraction of any kind. You are given N images of cats and M images of cars, with known annotations. Suggest an objective function, and explain how you learn the classifier. Are there any pre-requisites regarding the input?

[1] We did not talk about this in class, and it will not appear in the final exam. Please refer to `https://en.wikipedia.org/wiki/Bag-of-words_model_in_computer_vision`

Solution: Of course you can use support-vector machine. Here I would like to give another approach which finds a linear transformation where the two classes would be separable. The objective function aims to maximize the between class scatter while minimizing the inner class scatter.

The objective function is:

$$\max_{\boldsymbol{w}} \frac{\sum_{i=1}^{N_s}\sum_{j=1}^{N_m}(\boldsymbol{w}^T\boldsymbol{s}_i - \boldsymbol{w}^T\boldsymbol{m}_j)^2}{\sum_{i=1}^{N_s}\sum_{j=1}^{N_s}(\boldsymbol{w}^T\boldsymbol{s}_i - \boldsymbol{w}^T\boldsymbol{s}_j)^2 + \sum_{i=1}^{N_m}\sum_{j=1}^{N_m}(\boldsymbol{w}^T\boldsymbol{m}_i - \boldsymbol{w}^T\boldsymbol{m}_j)^2}$$

This is identical to

$$\max_{\boldsymbol{w}} \frac{\boldsymbol{w}^T\big(\sum_{i=1}^{N_s}\sum_{j=1}^{N_m}(\boldsymbol{s}_i - \boldsymbol{m}_j)(\boldsymbol{s}_i - \boldsymbol{m}_j)^T\big)\boldsymbol{w}}{\boldsymbol{w}^T\big(\sum_{i=1}^{N_s}\sum_{j=1}^{N_s}(\boldsymbol{s}_i - \boldsymbol{s}_j)(\boldsymbol{s}_i - \boldsymbol{s}_j)^T + \sum_{i=1}^{N_m}\sum_{j=1}^{N_m}(\boldsymbol{m}_i - \boldsymbol{m}_j)(\boldsymbol{m}_i - \boldsymbol{m}_j)^T\big)\boldsymbol{w}}$$

which admits the form of $\frac{\boldsymbol{w}^T A\boldsymbol{w}}{\boldsymbol{w}^T B\boldsymbol{w}}$, and the optimal value of $\boldsymbol{w}$ is given by solving the generalized eigenvector problem $A\boldsymbol{w} = \lambda B\boldsymbol{w}$.

Where $\boldsymbol{s}_i$ and $\boldsymbol{m}_j$ are the cats and cars images respectively. The solution is acquired through a generalized eigen-decomposition.Note that the images should be aligned.

12. (10 points) Linear Filter: A common approach to object detection consists of the following stages:

1. Generate object proposals (sub-images)
2. Each proposal is then divided into cells. Calculate histogram of the gradient directions, weighted by the gradient magnitude in each cell and concatenating the histograms into feature vector X.
3. Classify the vector X using a classifier of choice.

1. Discuss briefly the different factors, which influence the input image and may interfere with the detection. Briefly explain each one, and how this algorithm reduces the interference.
Solution:

- The image of the same object changes with illumination–the algorithm is based on gradient direction which is less sensitive to illumination than the intensity.
- The image of the same object changes with pose–the algorithm uses cell which contain parts, which change less. It uses histograms which are less sensitive to small pose changes. It uses a (learning based) classifier that can use examples of different poses.
- The image changes with changes within the class -the algorithm uses a (learning based) classifier that use different examples from the class.

2. For an input image of size N pixels, there are M object proposals, K cells in a proposal and each histogram consists of B bins. Give an estimation of the detection algorithm computational complexity. You can assume that the classifier is linear. When calculating the complexity, do so for the worst case.
Solution: Gradient calculation $O(N)$. For each proposal: Histograms construction $O(N)$–(with worst case assumption that each of the proposals is nearly as large as the image.). Constructing X from histograms– $O(KB)$. Classification with linear classifier $O(KB)$. Overall $O(N + KB)$. Overall: $O(N + M(N + KB)) = O(M(N + KB))$.
3. Can you suggest a method of lower computational complexity that implements the same detection process? Explain. Note:assume M is very large.
Solution: The trick is to use integral images:

- First find gradients, and quantize them into B bins, and for each quantized value, separately, build an image containing in each pixel the gradient size value (if the gradient direction corresponds to this quantized value) or 0 (otherwise). For each such image build an integral image. This takes $O(NB)$ time.
- For each cell build gradient direction histogram from integral image in $O(B)$ time. Build the vector for a proposal and classify in $O(KB)$ time.
- For all proposals, the overall complexity is $O(NB+MKB)$. As M is large(st) and N is the 2nd largest, this is better than the previous algorithmâs complexity.

13.(10 points) Image Classification: Let the following matrix H denote a general 2D planar transformation:

$$H = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} \tag{1}$$

For each of the following special cases, write down (i) the parameterization of the h_{ij}'s, (ii) the number of degrees of freedom, and (iii) the number of correspondences needed to estimate H.

- (a) Euclidean. **Solution:** three parameters θ, t_x, t_y:

$$H = \begin{pmatrix} \cos(\theta) & -\sin(\theta) & t_x \\ \sin(\theta) & \cos(\theta) & t_y \\ 0 & 0 & 1 \end{pmatrix}. \tag{2}$$

- (b) Similarity. **Solution:** four parameters a, b, t_x, t_y:

$$H = \begin{pmatrix} a & -b & t_x \\ b & a & t_y \\ 0 & 0 & 1 \end{pmatrix}. \tag{3}$$

- (c) Affine. **Solution:** six parameters:

$$H = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ 0 & 0 & 1 \end{pmatrix}. \tag{4}$$

- (d) Projective. **Solution:** eight parameters:

$$H = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{pmatrix}. \tag{5}$$

14.(10 points) K-means clustering: Given a dataset that consists of the following four points

$$\boldsymbol{x}_1 = (-1, 1),\ \boldsymbol{x}_2 = (1, 1),\ \boldsymbol{x}_3 = (-1, -1),\ \boldsymbol{x}_4 = (1, -1). \tag{6}$$

We want to do K-means clustering using Euclidean distances when $K = 2$. We start by randomly picking two points as the cluster centroids.
1. (a) What are all possible clustering results?
Solution:

$$\text{Cluster1} : \boldsymbol{x}_1, \boldsymbol{x}_2 \quad \text{Cluster2} : \boldsymbol{x}_1, \boldsymbol{x}_2 \tag{7}$$
$$\text{Cluster1} : \boldsymbol{x}_1, \boldsymbol{x}_3 \quad \text{Cluster2} : \boldsymbol{x}_2, \boldsymbol{x}_4 \tag{8}$$
$$\text{Cluster1} : \boldsymbol{x}_1 \quad \text{Cluster2} : \boldsymbol{x}_2, \boldsymbol{x}_3, \boldsymbol{x}_4 \tag{9}$$
$$\text{Cluster1} : \boldsymbol{x}_2 \quad \text{Cluster2} : \boldsymbol{x}_1, \boldsymbol{x}_3, \boldsymbol{x}_4 \tag{10}$$
$$\text{Cluster1} : \boldsymbol{x}_3 \quad \text{Cluster2} : \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_4 \tag{11}$$
$$\text{Cluster1} : \boldsymbol{x}_4 \quad \text{Cluster2} : \boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3 \tag{12}$$

2. (b) Among all possible clustering results, which have the smallest cost, as measured in total distance (defined as follows). We are given points $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_4 \in R^n$ and we wish to organize these points into 2 clusters. This amounts to choosing a level for each point, where each label has two choices. The total distance is defined as $\sum_{i=1}^{4} d(\boldsymbol{x}_i, \mu_i)$ where μ_c is the center of all points $\boldsymbol{x}_i$ with the same label c, and $d(\boldsymbol{x}, \boldsymbol{y})$ is the Euclidean distance between the points $\boldsymbol{x}$ and $\boldsymbol{y}$.
Solution: The last four clustering results have the smallest cost.
3. In a situation like above, how could you modify the K-means algorithm to output a clustering result that has relatively small total distances (defined above)?
Solution: Try multiple random initializations, and choose the result with the smallest total cost to output.

3 Short Answer

15. **(5 points)** What differentiate intrinsic and extrinsic camera parameters.
Solution: Intrinsic camera parameters such as focal length do not change when varying the camera poses. Extrinsic camera parameters are dependent on camera poses.

16. **(5 points)** Given two images, what is the difference between structure-from-motion and stereo matching?
Solution: Structure-from-motion estimate intrinsic and extrinsic camera parameters, which characterize the search space

17. **(5 points)** Besides the uniqueness constraint, what are two other objectives that are used in stereo matching?
Solution: Smoothness and ordering constraints.

18. **(5 points)** How many feature points do we need to solve the image calibration problem?
Solution: 6 points and they should not lie on the same plane.

19. **(6 points) True / False : Essential Matrix**

- (2 points) What is the rank of the essential matrix?

  Solution:

  The rank is 2.

- (4 points) In the 8-point algorithm, what math technique is used to enforce the estimated fundamental matrix to have the proper rank? Explain how this math technique is used to enforce the proper matrix rank.

  Solution:

  In the 8-point algorithm, SVD can be used to enforce the estimated E has rank 2. Specifically, we compute the SVD decomposition $E = U\Sigma V$, and we then zero out diagnals of Σ and average the two largest singular values to obtain $\Sigma = \text{diag}(\frac{\sigma_1+\sigma_2}{2}, \frac{\sigma_1+\sigma_2}{2}, 0)$. We can reconstruct $E = U\Sigma V$.

20. **(3 points)** What is the basic task for visual recognition? Please give a concrete example.
Solution: Image classification is the basic task. For example, in object detection, we can do sliding window and classify the image patches.

21. **(3 points)** How to improve the performance of sliding window? List one technique
Solution: Early stop using more efficient classifiers.

22. **(3 points)** What is the key advantage of KNN classifier?
Solution: Works well if the size of the training dataset is large enough.

23. **(3 points)** What is the key advantage of SVM classifier versus KNN classifier?
Solution: It is more efficient and works better with modest training datasets.