Q1 You are given the facial images of 100 persons. Each person has N images, each with a size of 360×260 pixels. For the k-th person, his/her images are denoted as $\mathcal{X}_k = \{\mathbf{x}_{k,1}, \ldots, \mathbf{x}_{k,N}\}$, where $\mathbf{x}_{k,i}$ is a vector formed by stacking the columns of the corresponding image. Assume that the images of these 100 persons share a global covariance matrix:

$$\boldsymbol{\Sigma} = \frac{1}{100N} \sum_{k=1}^{100} \sum_{i=1}^{N} (\mathbf{x}_{k,i} - \boldsymbol{\mu})(\mathbf{x}_{k,i} - \boldsymbol{\mu})^{\mathsf{T}},$$

where

$$\boldsymbol{\mu} = \frac{1}{100N} \sum_{k=1}^{100} \sum_{i=1}^{N} \mathbf{x}_{k,i}.$$

Assume also that a Gaussian classifier based on $\boldsymbol{\Sigma}$ is used for classifying these 100 persons.

(a) What is the theoretical minimum value of N for the Gaussian classifier to recognize these 100 persons?

(5 marks)

(b) Determine the theoretical minimum value of N if we have the following relationship:

$$\mathbf{x}_{k,2} = \mathbf{x}_{k,1} \odot \mathbf{x}_{k,1} + 2\mathbf{x}_{k,1} + \mathbf{1} \quad \text{for all} \quad k = 1, \ldots, 100,$$

where $\odot$ denotes the element-wise products of vectors and $\mathbf{1}$ is a vector containing all 1's. Briefly explain your answer.

(5 marks)

(c) Briefly describe the procedure for training the Gaussian classifier. Write the equation that computes the mean vectors of the Gaussian classifier.

(5 marks)

(d) Are the decision boundaries produced by this Gaussian classifier linear or nonlinear? Briefly explain your answer.

(5 marks)

(e) If a deep neural network is used for classifying these 100 persons, how would you make sure that the outputs of the network give the posterior probabilities of these persons given an unknown image, i.e., $P(\text{Person} = k|\mathbf{x})$ for all $k = 1, \ldots, 100$? Your answer should specify the number of input nodes, the number of output nodes, and the type of output nodes.

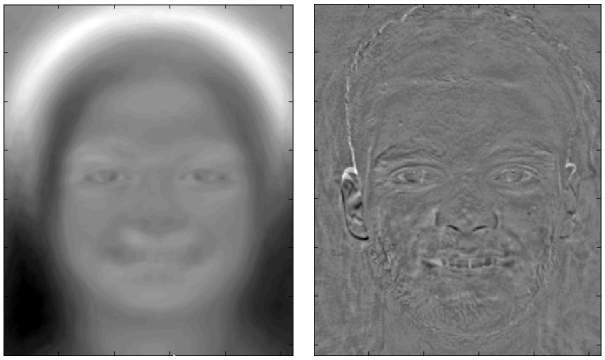
(5 marks)

Q2 Assume that you have a training set comprising the facial images of N males and N females. Each image has a size of 64×64 pixels. The male images are denoted as $\mathcal{X}^{(m)} = \{\mathbf{x}_1^{(m)}, \ldots, \mathbf{x}_N^{(m)}\}$, where $\mathbf{x}_i^{(m)}$ is a vector formed by stacking the columns of the i-th male image. Similarly, the female images are denoted as $\mathcal{X}^{(f)} = \{\mathbf{x}_1^{(f)}, \ldots, \mathbf{x}_N^{(f)}\}$. Also, denote $\mathcal{X} = \{\mathcal{X}^{(m)}, \mathcal{X}^{(f)}\}$ as the whole training set. To build a gender recognition system, you first reduce the dimension of the vectors in $\mathcal{X}$ and then apply the low-dimensional vectors to train a support vector machine (SVM) classifier.

(a) How would you reduce the dimension of the vectors in $\mathcal{X}$ to 5 using principal component analysis? You may assume that your computer has a large amount of memory and computation time is not an issue.

(6 marks)

(b) Fig. Q2 shows the eigenfaces corresponding to the first principal component (PC) and the eigenface corresponding to the last PC with the smallest positive eigenvalue. Identify the image [(a) or (b)] corresponding to the last PC. Briefly explain your answer. (4 marks)



(a) (b)

Fig. Q2

(c) If $N = 5000$, what is the maximum number of eigenfaces (with positive eigenvalues) that you could obtain? Briefly explain your answer.

(5 marks)

(d) What is the number of eigenfaces with positive eigenvalues if linear discriminant analysis (LDA) is used for dimension reduction? You may assume that the LDA uses the genders as the class labels. Briefly explain your answer.

(5 marks)

(e) Assume that the numbers of male and female images are 1 and 100, respectively. Without dimension reduction, suggest the most appropriate kernel type (linear, polynomial or RBF) for the SVMs. Briefly explain your answer.

(5 marks)

Q3 You are requested to develop a hidden Markov model (HMM) based speech recognizer that can recognize the words "Yes" and "No".

(a) Assuming that the occurrences of "Yes" and "No" have equal prior probability, draw a block diagram to illustrate the structure of the recognizer. Your diagram should contain blocks depicting the feature extractor, HMMs, and decision logic.

(5 marks)

(b) Outline the procedure for training the HMMs in this recognizer.

(6 marks)

(c) Given that both "Yes" and "No" comprise three phonemes, suggest the number of states in the HMMs. Briefly explain your answer.

(4 marks)

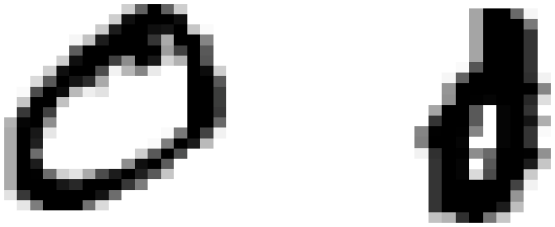
(d) Denote the likelihoods of an acoustic sequence $\mathcal{X}$ as $p(\mathcal{X}|\Lambda_{\text{yes}})$ and $p(\mathcal{X}|\Lambda_{\text{no}})$, where Λ_{yes} and Λ_{no} are the HMM of the words "Yes" and "No", respectively. Also denote the prior probabilities for "Yes" and "No" as $P(\text{'Yes'})$ and $P(\text{'No'})$, respectively. If $P(\text{'Yes'}) = 0.2$, explain how you would use the likelihoods and the prior probabilities to classify $\mathcal{X}$. Your answer should contain an expression relating the predicted word, the prior probabilities, and the likelihoods.

(6 marks)

(e) Are pronunciation dictionaries and language models necessary for this recognizer? Briefly explain your answer.

(4 marks)

Q4 (a) A training set comprising 10,000 images was used to train a handwritten digit recognizer based on support vector machines (SVMs). Each of the digits ('0' to '9') has 1,000 images of size 28×28 pixels. The recognizer comprises 10 one-vs-rest SVMs, each responsible for classifying one digit against the remaining digits. Fig. Q4 shows two of the Digit '0'; they are part of the training set for training the SVM responsible for classifying Digit '0' against Digits '1'–'9'.



(a) (b)

Fig. Q4

(i) Which of the images [(a) or (b)] in Fig. Q4 corresponds to a support vector? Briefly explain your answer.

(6 marks)

(ii) Assume that the SVMs in the recognizer use a radial basis function (RBF) as their kernel, i.e.,

$$K(\mathbf{x}, \mathbf{x}_i) = \exp\left\{-\frac{\|\mathbf{x} - \mathbf{x}_i\|^2}{2\sigma^2}\right\},$$

where σ is the kernel parameter and $\mathbf{x}_i$'s are support vectors. What will be the theoretical maximum number of support vectors in each SVM? Suggest the value of σ that will cause the SVMs to have the maximum number of support vectors.

(5 marks)

(iii) If the number of training images reduces to one per digit and linear SVMs are used in the recognizer, what will be the minimum number of support vectors for each SVM?

(3 marks)

(b) In GMM-UBM and GMM-SVM speaker verification, given a sequence of acoustic vectors $\mathcal{X}^{(s)}$ from a client speaker s, the maximum a posteriori (MAP) adaptation is used for adapting the universal background model (UBM) to create the speaker-dependent Gaussian mixture model (GMM). Typically, only the mean vectors of the UBM are adapted:

$$\boldsymbol{\mu}_j^{(s)} = \alpha_j E_j(\mathcal{X}^{(s)}) + (1 - \alpha_j)\boldsymbol{\mu}_j^{\text{ubm}}, \qquad j = 1, \ldots, M,$$

where M is the number of mixture components in the UBM, $E_j(\mathcal{X}^{(s)})$ is the sufficient statistics depending on $\mathcal{X}^{(s)}$, and $\boldsymbol{\mu}_j^{\text{ubm}}$ and $\boldsymbol{\mu}_j^{(s)}$ are the j-th mean vector of the UBM and the adapted GMM, respectively.

(i) Discuss the value of α_j when the enrollment utterance is very long and when the enrollment utterance is very short.

(6 marks)

(ii) In GMM-SVM speaker verification, we stack the mean vectors $\boldsymbol{\mu}_j^{(s)}$ for $j = 1, \ldots, M$ to construct a speaker-dependent supervector vector:

$$\vec{\boldsymbol{\mu}}^{(s)} = \left[\left(\boldsymbol{\mu}_1^{(s)}\right)^\mathsf{T} \ \cdots \ \left(\boldsymbol{\mu}_M^{(s)}\right)^\mathsf{T} \right]^\mathsf{T}.$$

Why is it important to use MAP instead of directly applying the EM algorithm to compute $\boldsymbol{\mu}_j^{(s)}$'s when constructing $\vec{\boldsymbol{\mu}}^{(s)}$?

(5 marks)

– END OF PAPER –