

Midterm Exam Solutions

Prof. Jim Rehg
CS 7635 Computational Perception
College of Computing
Georgia Institute of Technology

April 2, 2002

1. Face Recognition [25 points]

[Note: This is a subjective question, so we accepted any reasonably-argued answer.]

1. Pose variation and lighting variation. Most face recognition systems do not work well when pose variations exceed 30 degrees, and most cannot handle extreme changes in illumination. This is a key issue because it is tied directly to the role of 3-D shape in recognition, which is an important unresolved issue in object recognition generally. Pose and illumination variations are also important in practical applications. Of the two factors, pose may be the bigger problem simply because it is harder to control it by engineering the task environment. But the two are interrelated through their dependence on geometry.

2. Variations in appearance over short to long time scales: appearance changes due to facial expressions, hairstyle, glasses, blemishes, etc. The FERET study documented the deleterious effects of variation in appearance on recognizer performance. In contrast, while people may not be able to explain precisely what has changed in someone's appearance, they are remarkably good at recognition in the face of significant appearance changes.

3. Recognition using cues from the head/torso. Most face recognition systems use a heavily-cropped image of the face which preserves the basic facial features but discards information about the hair, hairline, neckline, shoulders, etc. People can make use of these other cues when they are available, and so should recognizers. There are similarities to the problem of recognition from profile views. In both cases it is no longer possible to define a window of pixels that is guaranteed to contain only pixels from the face class, assuming that a face is present. It seems important to push from this "window-oriented" approach to one that can handle the implicit segmentation problem.

One promising avenue for future work is to look at methods which describe a face as a combination of learned image features. The feature values encode intensity information, and their spatial configuration encodes geometric information. A collection of features used in this manner has more flexibility than a single template model. Recent work in face detection by Leung and Perona, Viola and Jones, and Schneiderman and Kanade fall into this category. How to extend these ideas to face recognition is an open question.

2. PCA [25 points]

We are given:

- (1) $\mathbf{Y}\mathbf{Y}^T \mathbf{u}_i = \sigma_i^2 \mathbf{u}_i$
- (2) $\mathbf{Y}^T \mathbf{Y} \mathbf{v}_i = \lambda_i^2 \mathbf{v}_i$

(a) $\|\mathbf{Y}^T \mathbf{u}_i\| = \sqrt{\mathbf{u}_i^T \mathbf{Y}\mathbf{Y}^T \mathbf{u}_i} = \sigma_i$ (since the \mathbf{u}_i are orthonormal). Similarly, $\|\mathbf{Y} \mathbf{v}_i\| = \lambda_i$.

(b) Both $\mathbf{Y}\mathbf{Y}^T$ and $\mathbf{Y}^T \mathbf{Y}$ are symmetric and positive semidefinite. Therefore, they both have a full set of orthogonal eigenvectors and real, nonnegative eigenvalues.¹ Left multiplying Equations (1) and (2) by \mathbf{Y}^T and \mathbf{Y} , respectively, and grouping terms yields:

$$\begin{aligned}\mathbf{Y}^T \mathbf{Y} (\mathbf{Y}^T \mathbf{u}_i) &= \sigma_i^2 (\mathbf{Y}^T \mathbf{u}_i) \\ \mathbf{Y}\mathbf{Y}^T (\mathbf{Y} \mathbf{v}_i) &= \lambda_i^2 (\mathbf{Y} \mathbf{v}_i)\end{aligned}$$

Suppose $\sigma_i^2 \neq 0$. We must have $\mathbf{Y}^T \mathbf{u}_i \neq \mathbf{0}$ since $\mathbf{Y}\mathbf{Y}^T \mathbf{u}_i \neq \mathbf{0}$. It follows that $(\mathbf{Y}^T \mathbf{u}_i, \sigma_i^2)$ are an eigenvector/eigenvalue of $\mathbf{Y}^T \mathbf{Y}$. Similarly, when $\lambda_i^2 \neq 0$ we have that $(\mathbf{Y} \mathbf{v}_i, \lambda_i^2)$ are an eigenvector/eigenvalue of $\mathbf{Y}\mathbf{Y}^T$. It follows that $\mathbf{Y}\mathbf{Y}^T$ and $\mathbf{Y}^T \mathbf{Y}$ share the same set of nonzero eigenvalues, and there can be at most p of them since $\text{rank}(\mathbf{Y}^T \mathbf{Y}) \leq p$. We have established that $\sigma_i = \lambda_i$ (the equality is trivial in case where the eigenvalues are zero).

For each nonzero σ_i^2 , the vector $\hat{\mathbf{v}}_i = \mathbf{Y}^T \mathbf{u}_i$ is an eigenvector of $\mathbf{Y}^T \mathbf{Y}$ associated with λ_i^2 . It follows that

$$\mathbf{v}_i = \frac{\hat{\mathbf{v}}_i}{\|\hat{\mathbf{v}}_i\|} = \frac{\mathbf{Y}^T \mathbf{u}_i}{\sigma_i},$$

where the last equality comes from part (a). Similarly, we can obtain

$$\mathbf{u}_i = \frac{\mathbf{Y} \mathbf{v}_i}{\lambda_i}.$$

¹To show that a symmetric matrix \mathbf{A} has orthogonal eigenvectors, let $\mathbf{A}\mathbf{x}_1 = \lambda_1 \mathbf{x}_1$ and $\mathbf{A}\mathbf{x}_2 = \lambda_2 \mathbf{x}_2$, where $\lambda_1 \neq \lambda_2$. Then $\mathbf{x}_1^T \mathbf{A}\mathbf{x}_2 = \lambda_1 \mathbf{x}_1^T \mathbf{x}_2 = \lambda_2 \mathbf{x}_1^T \mathbf{x}_2$. Subtracting yields $(\lambda_1 - \lambda_2) \mathbf{x}_1^T \mathbf{x}_2 = 0$ which implies $\mathbf{x}_1^T \mathbf{x}_2 = 0$. In the repeated eigenvalue case there is still a full set of linearly independent eigenvectors, and they can be made orthogonal by Gram-Schmidt.

By dividing Equations (1) and (2) by σ_i and λ_i , respectively, and substituting the previous relations, we obtain:

$$(3) \quad \frac{\mathbf{Y}(\mathbf{Y}^T \mathbf{u}_i)}{\sigma_i} = \frac{\sigma_i^2 \mathbf{u}_i}{\sigma_i} \Rightarrow \mathbf{Y} \mathbf{v}_i = \sigma_i \mathbf{u}_i.$$

$$(4) \quad \frac{\mathbf{Y}^T(\mathbf{Y} \mathbf{v}_i)}{\lambda_i} = \frac{\lambda_i^2 \mathbf{v}_i}{\lambda_i} \Rightarrow \mathbf{Y}^T \mathbf{u}_i = \lambda_i \mathbf{v}_i.$$

Now consider the eigenvectors in the zero eigenvalue case. Let r be the rank of \mathbf{Y} , with $r \leq p < n$. Let $\bar{\mathbf{v}}_i$ be one of the $p - r$ eigenvectors of $\mathbf{Y}^T \mathbf{Y}$ with a zero eigenvalue. From part (a) we have $\|\mathbf{Y} \bar{\mathbf{v}}_i\| = 0 \Rightarrow \mathbf{Y} \bar{\mathbf{v}}_i = \mathbf{0}$. Since $\bar{\mathbf{v}}_i \neq \mathbf{0}$, it follows that the $\bar{\mathbf{v}}_i$ form a basis for the (right) nullspace of \mathbf{Y} . Similarly, we have that the $n - r$ eigenvectors $\bar{\mathbf{u}}_i$ satisfy $\mathbf{Y}^T \bar{\mathbf{u}}_i = \mathbf{0}$ and form a basis for the left nullspace of \mathbf{Y} . This establishes the validity of Equations (3) and (4) in the zero eigenvalue case.

(c) The principal components of \mathbf{S} are the left singular vectors of the data matrix \mathbf{Y} (with the means already subtracted out):²

$$\mathbf{S} = \frac{1}{N} \sum_{i=1}^N \mathbf{y}_i \mathbf{y}_i^T = \frac{1}{N} \mathbf{Y} \mathbf{Y}^T$$

$$\mathbf{S} \mathbf{u}_i = \alpha_i \mathbf{u}_i$$

$$\alpha_i = \frac{\sigma_i^2}{N}$$

The complexity of the eigenvalue problem for \mathbf{S} (using an algorithm like symmetric QR) is $O(n^3)$. In contrast, the right singular vectors $\{\mathbf{v}_i\}$ and singular values $\{\lambda_i^2\}$ can be computed in $O(p^3)$, a dramatic savings for $p \ll n$. The principal components can then be obtained as follows:

$$\alpha_i = \frac{\lambda_i^2}{N}$$

$$\mathbf{u}_i = \frac{\mathbf{Y} \mathbf{v}_i}{\lambda_i}$$

²Although many authors have rediscovered the SVD trick, the first reference seems to be H. Murakami and V. Kumar, "Efficient Calculation of Primary Images from a Set of Images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (4)5:511-515, September 1982.

(d) The connection to the SVD comes from collecting the instances of Equation 3 into a matrix equation:

$$\begin{aligned}
 (5) \quad \mathbf{Y} [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \cdots \quad \mathbf{v}_p] &= [\sigma_1 \mathbf{u}_1 \quad \sigma_2 \mathbf{u}_2 \quad \cdots \quad \sigma_p \mathbf{u}_p] \\
 (6) \quad &= [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \cdots \quad \mathbf{u}_n] \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \sigma_2 & & \vdots & & \vdots \\ \vdots & & \ddots & 0 & & \vdots \\ 0 & \cdots & 0 & \sigma_r & 0 & \cdots & 0 \\ \vdots & & & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \end{bmatrix}
 \end{aligned}$$

$$(7) \quad \mathbf{YV} = \mathbf{U}\Sigma$$

where \mathbf{V} is the p by p orthonormal matrix with columns $\{\mathbf{v}_i\}$, \mathbf{U} is the n by n orthonormal matrix with columns $\{\mathbf{u}_i\}$, and Σ is the n by p matrix which is zero everywhere except for the first r diagonal entries: $[\Sigma]_{ii} = \sigma_i, 1 \leq i \leq r$ (the remaining $p - r$ singular values are zero). Since $\mathbf{V}^{-1} = \mathbf{V}^T$, Equation (7) yields the familiar SVD equation: $\mathbf{Y} = \mathbf{U}\Sigma\mathbf{V}^T$.

3. Markov Chains [15 points]

[Note: perhaps the most straightforward way to solve this problem is to sketch the probabilistic state machine and demonstrate, by visual inspection of the paths, that some states are unreachable from others. If you solved it this way you got full credit.]

Construct Π so that the i th column is the distribution $p(s_t | s_{t-1} = i)$. Then the j th row is the set $p(s_t = j | s_{t-1})$. From $p(s_t = j) = \sum_{i=1}^n p(s_t = j | s_{t-1} = i)p(s_{t-1} = i)$ we have $p(s_t) = \Pi p(s_{t-1})$. Recursive expansion yields $p(s_t) = \Pi^t p(s_0)$ and $\Pi^t(i, j) = p(s_t = i | s_0 = j)$. In general we have $p(s_{t_2}) = \Pi^{(t_2-t_1)} p(s_{t_1})$.³ If the Markov model is to be irreducible, there must be some value for t such that Π^t has no zero elements. Then, given any $\tau \geq t$, Π^τ also has all positive elements, since it consists of inner products between positive and nonnegative vectors.

$$(1) \quad \Pi = \begin{bmatrix} 0.5 & 0.9 & 0.0 & 0.1 \\ 0.5 & 0.1 & 0.2 & 0.0 \\ 0.0 & 0.0 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.3 & 0.9 \end{bmatrix}$$

Since the upper block triangular structure of a matrix is preserved by repeated multiplications with itself, this markov chain is not irreducible (the block of zeroes in the lower left corner will never go away). States 3 and 4 form a *transient set*. Once the state machine leaves the transient set it will never return. States 1 and 2 form an *ergodic set*. Once the state machine enters the ergodic set it will never leave.

$$(2) \quad \Pi = \begin{bmatrix} 0.0 & 0.0 & 1.0 & 0.0 \\ 0.8 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.2 & 0.0 & 1.0 \\ 2.0 & 0.8 & 0.0 & 0.0 \end{bmatrix}, \quad \Pi^9 = \begin{bmatrix} 10.0 & 4.3 & 0.41 & 2.76 \\ 0.33 & 0.75 & 2.2 & 3.7 \\ 9.0 & 2.3 & 10.0 & 0.41 \\ 2.6 & 2.2 & 8.5 & 9.3 \end{bmatrix}$$

The markov chain is irreducible.

$$(3) \quad \Pi = \begin{bmatrix} 0.25 & 0.0 & 0.0 & 0.0 \\ 0.25 & 0.0 & 0.0 & 1.0 \\ 0.25 & 1.0 & 0.0 & 0.0 \\ 0.25 & 0.0 & 1.0 & 0.0 \end{bmatrix}$$

³Note that the matrices in the problem were given in the transposed form, so that the rows rather than columns summed to one (a legacy of the convention that I used in class). You can either transpose them or multiply on the left rather than on the right.

The lower block triangular structure is preserved under repeated multiplication, and as a consequence the three zeros in the first row will never go away. The Markov chain is not irreducible. The transient set is $\{1\}$. The ergodic set is $\{2, 3, 4\}$.

4. Markov Models [15 points]

(a) Let the states be: d - dropped out, g - graduated, and $\{y_1, y_2, y_3, y_4\}$ - the four years of high school.⁴ Let the state vector be $s = [d \ g \ y_4 \ y_3 \ y_2 \ y_1]$. We then have $\pi_0 = [0 \ 0 \ 0 \ 0 \ 0 \ 1]$ and

$$\Pi = \begin{bmatrix} 1 & 0 & d & d & d & d \\ 0 & 1 & a & 0 & 0 & 0 \\ 0 & 0 & r & a & 0 & 0 \\ 0 & 0 & 0 & r & a & 0 \\ 0 & 0 & 0 & 0 & r & a \\ 0 & 0 & 0 & 0 & 0 & r \end{bmatrix}.$$

This is a valid state transition matrix provided $d + a + r = 1$.

(b) From problem 3 we have $p(s_t) = \Pi p(s_{t-1})$. Since the problem asks to count years starting at 1, we have

$$p(s_3) = \Pi^2 \pi_0 = \begin{bmatrix} d(1 + a + r) \\ 0 \\ 0 \\ a^2 \\ 2ra \\ r^2 \end{bmatrix}, \text{ and } p(s_3 = y_3) = a^2.$$

⁴This problem and its solution are taken from the book *Finite Markov Chains* by John G. Kemeny and J. Laurie Snell, Van Nostrand, 1960, p. 30. This is a standard reference for the material in questions 3 and 4.

5. Hidden Markov Models [20 points]

(a) From the definitions:

$$\begin{aligned} p(\mathbf{Y}_T) &= \sum_{k=1}^N p(\mathbf{Y}_T, s_T = k) = \sum_{k=1}^N \alpha_T(k) \\ p(\mathbf{Y}_T) &= \sum_{k=1}^N p(\mathbf{Y}_T | s_0 = k) p(s_0 = k) = \sum_{k=1}^N \beta_0(k) \pi_0(k) \end{aligned}$$

(b) From the definitions:

$$\begin{aligned} \gamma_t(k) &= \frac{p(s_t = k, \mathbf{Y}_t, \bar{\mathbf{Y}}_{t+1})}{p(\mathbf{Y}_T)} \\ &= \frac{p(\bar{\mathbf{Y}}_{t+1} | \mathbf{Y}_t, s_t = k) p(\mathbf{Y}_t, s_t = k)}{p(\mathbf{Y}_T)} \\ &= \frac{\alpha_t(k) \beta_t(k)}{\sum_{j=1}^n \alpha_t(j) \beta_t(j)} \end{aligned}$$