Final exam: CS 663, Digital Image Processing, 20$^{\text{th}}$ November

Instructions: There are 180 minutes for this exam (2 pm to 5 pm). Answer all questions. This exam is worth 20% of the final grade. Some formulae are listed at the end of the paper. You may re-use results used in class directly. **Write brief answers. Lengthy answers are not required.**

1. Briefly explain the advantages of the YCbCr color scheme over the RGB color scheme in image compression. [10 points]

   Solution: The YCbCr color space is a decorrelated color space as it is related to PCA, and hence the Y, Cb, Cr values are not correlated with each other unlike the R,G,B values. As a result, one can independently code the Y, Cb, Cr channels using the JPEG method designed for grayscale images without causing color artifacts despite quantization. Moreover, the Cb and Cr channels can be downsampled by some factor (usually 2) without causing noticeable artifacts as the human visual system is less sensitive to changes in chrominance (Cb and Cr) as compared to Y. This cannot be done with RGB.
   Marking Scheme: 5 points for decorrelation and 5 points for downsampling of chrominance.

2. Suppose you are at the Taj Mahal and wish to take a picture of the monument. But there are people moving around and they would show up in the picture, whereas you wish to have a picture of the monument without such occlusions. Suppose you have access to a video camera with a tripod stand. Given your knowledge of image processing, how will you produce such a picture from a video sequence containing occlusions? Will there be any change in your strategy for color (RGB) images versus grayscale images? Explain. For everything, assume that (1) there is no illumination variation during acquisition of the video sequence, (2) no motion blur, and (3) that no pixel contains an occlusion for all frames. [10 points]

   Solution: The solution is a pixelwise median filter across time. For grayscale images, there will be three independent pixelwise median filters across time - one each for R,G,B. There is an innate assumption that there is no pixel that is occluded for more than 50% of the time instants.
   Marking scheme: 7 points for median filter solution for grayscale images and 3 points for the extension to color images. The 50% assumption need not be stated. For mean filter, deduct 5 points out of 10, as the mean filter is not robust.

3. (a) Consider a matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$. Explain how you will find matrix $\boldsymbol{B}$ of rank $r < \min(m, n)$ such that $\sum_{i=1}^{m} \sum_{j=1}^{n} (A_{ij} - B_{ij})^2$ is the least possible.

   Solution: Compute the SVD of $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$. The required approximation $\boldsymbol{B} = \boldsymbol{U_r}\boldsymbol{S_r}\boldsymbol{V_r}^T$ where S_r is a $m \times r$ diagonal matrix containing the top r singular values of $\boldsymbol{A}$ and $\boldsymbol{U_r}, \boldsymbol{V_r}$ represent the corresponding r left and right singular vectors respectively. As per the Eckart-Young theorem, this is guaranteed to produce the approximation with least error as defined in the question.

   (b) Why is the DCT preferred over the DFT for image compression in standards such as JPEG? [6+6=12 points]

   Solution: DCT is preferred because it provides better energy compact as it is obtained from the DFT of a symmetrically padded image, which avoid the border discontinuities typical for DFT. It is also closely related to PCA for natural image patches or any stationary first order Markov process with correlation ρ close to 1.
   Marking scheme: Deduct 2 points if there is no mention of PCA. Deduct 2 points if a student states that it is rela-valued and hence requires less storage. DFT for a real image is symmetric and hence that advantage does not exist.

4. Let $\boldsymbol{M} \in \mathbb{R}^{N \times Q}$ be a matrix, each column of which contains the image of a person's face (image in vectorized form with N pixels), but under one out of Q distinct lighting directions, assuming a distant point light source. Each image belongs to the same person and under the same pose. Prove that $\boldsymbol{M}$ has rank 3 under the Lambertian reflectance model which states that the intensity observed at a pixel is equal to $L \cos\theta$ where L is the intensity of the light source and θ is the angle between the light source direction and the surface normal at the object location which projects onto that pixel. What are the implications of this fact for face recognition under lighting variations? [12 points]

   Solution: Let $\boldsymbol{m_i}$ be the i^{th} column of $\boldsymbol{M}$. Then $\boldsymbol{m_i} = L\boldsymbol{Z}\boldsymbol{s_i}$ where $\boldsymbol{Z}$ is a $N \times 3$ matrix, each row of which contains the 1×3 surface normal vector at the physical point corresponding to each pixel, and where $\boldsymbol{s_i}$ is a 3×1 lighting direction vector. This tells us that $\boldsymbol{M} = L\boldsymbol{Z}\boldsymbol{S}$ where $\boldsymbol{S}$ is a $3 \times Q$ vector, each column of which contains one of Q lighting directions. Since $\boldsymbol{M}$ is a product of a $N \times 3$ matrix with a $3 \times Q$ matrix, it has rank 3. The implications of this for face recognition is that the set of images of a person's face with no viewpoint variation but with change in lighting direction forms a low-dimensional subspace. As lighting changes represent more variation that identity change, the top 3 eigenfaces represent illumination change. Hence, we can achieve illumination invariance in face recognition by removing the top three eigenfaces before comparing eigencoefficients.
   Marking scheme: 7 points for proof of rank 3, and 5 points for the implications for face recognition. Deduct 2 points if there is no mention that the top 3 eigenfaces account for illumination change since illumination represents more variation than identity.

5. Consider two $N \times N$ images $I_1(x, y)$ and $I_2(x, y)$ which are related to each other through an unknown translation t_x, t_y. Give an efficient algorithm to find t_x and t_y and state its time complexity. Also consider that v is a $N \times 1$ tomographic projection vector of image $I_1(x, y)$ taken at an *unknown* angle θ. Give an efficient algorithm to determine θ from v and I_1 and state its time complexity. [6+6=12 points]

   Solution: Let $\hat{I}_1$ and $\hat{I}_2$ represent the DFT of images I_1 and I_2 respectively. By fourier shift theorem, we know that $\hat{I}_2(u, v) = \hat{I}_1(u, v)e^{-j2\pi(t_x x + t_y y)/N}$. Hence we have $K(u, v) \triangleq \dfrac{\hat{I}_1^*(u, v)\hat{I}_2(u, v)}{|\hat{I}_1(u, v)||\hat{I}_2(u, v)|} = e^{j2\pi(t_x x + t_y y)/N}$. The inverse DFT of K is an impulse at (t_x, t_y). This algorithm takes time $O(N^2 \log N)$ for the DFT and IDFT computations and $O(N^2)$ time for computing K. The overall time complexity is $O(N^2 \log N)$ for an $N \times N$ image.

   For the second part, let $\hat{v}$ be the DFT of v. By Fourier slice theorem, it is equal to a slice through $\hat{I}_1$ at angle θ. Hence, we can trace out different slices in $\hat{I}_1$ at some Q angles using fourier-domain interpolation and compare with $\hat{v}$. The total time complexity is $O(N^2 \log N + NQ)$.
   Marking Scheme: For first part: Algorithm with principle 4 points (Deduct 2 points if there is no principle explained). 2 points for time complexity. Give only 2 out of 6 for a brute-force search algorithm which searches for t_x, t_y. For second part: 4 points for algorithm with the FST principle and 2 points for time complexity. For a brute-force algorithm that simply projects the image through different angles, give only 2 points out of 6. Such an algorithm has a complexity of $O(N^2 Q)$.

6. A color (RGB) image of our playground shows some portions covered with grass and some portions covered with mud. If you wished to segment such an image into two regions using the mean shift algorithm, on what features at each pixel will you build a probability density? Now consider a color image with three regions, each region containing a certain visibly distinct type of grass. The three regions have identical color (RGB) histograms. To segment this image using the mean shift algorithm, on what features at each pixel will you build a probability density? [6+6=12 points]

   Solution: For the first part, build a PDF on (x, y, R, G, B) at each pixel. For the second part, build your PDF on $(x, y, \boldsymbol{v})$ where $\boldsymbol{v}(x, y)$ contains various types of gradients (first order, second order in different directions and in different color channels) OR it contains the Fourier transform computed over a small window around (x, y) OR it contains the local structure tensor in that region in each channel. For the second part, any one correct answer will get full points.

7. Consider a planar object performing translational motion with constant velocity (a, b) (these are the X and Y

components of the velocity vector) in a plane exactly parallel to the plane of the CCD array of a stationary camera. The image of the object acquired by the camera during this motion will obviously be motion-blurred, and it will contain a small amount of noise, which we assume obeys a zero mean Gaussian distribution with small but unknown variance σ^2. Derive an expression for the motion blur kernel in the frequency domain and comment on the corresponding spatial domain kernel. Devise a semi-automated procedure to motion-deblur the image (acquired by a stationary camera) of a scene containing three similar planar objects, where one object is stationary and the other two are moving in a plane parallel to the CCD array of the camera with constant but unknown velocities (a_1, b_1) and (a_2, b_2) respectively. Assume that the three objects do not occlude or overlap with each other during the acquisition, and that the images are grayscale. Clearly state any other suitable assumptions you make. [4+4+8=16 points]

Solution: See lecture notes. The frequency domain expression is $H(u,v) = \dfrac{T\sin(\pi(au+bv))}{\pi(au+bv)}e^{-j\pi(au+bv)}$. The corresponding spatial domain filter consists of averaging of values in a nieghborhood around pixel (x, y) but constrained to lie along a direction (a, b) and constrained to lie in a rectangle of size $H \times W$ around the pixel where H pixels is equivalent to aT and W pixels is equivalent to a distance bT. We will use a Wiener filter approach to deblur the image. Mark out rectangular bounding boxes around the two moving objects to yield subimages I and J. Compute the Fourier transform of the two subimages. Then we will use the Wiener filter to the Fourier transform of the filtered image $\hat{I}_{filtered}(u,v) = \dfrac{H^*(u,v)I(u,v)}{|H(u,v)|^2 + S_n(u,v)/S_f(u,v)}$ where $H(u,v)$ is the frequency response of the blur kernel, $S_n(u,v)$ is the power spectrum of the noise and $S_f(u,v)$ is the power spectrum of the class of images to which I belongs. We can obtained $H(u,v)$ if we knew (a,b) which can be guessed by viewing the plot of $|I(u,v)|$, which will show a sinc-like pattern along the direction (a,b) in the frequency plane. Also $I(u,v) \approx 0$ (approximate because of small additive noise) whenever $au + bv$ is an integer. Hence the direction θ of (a,b) can be manually guessed by observing the plot. We can express $a = \rho\cos\theta, b = \rho\sin\theta$. We can find ρ by subsituting in the equation $\rho(u\cos\theta + v\sin\theta) = k$ for any integer k, since this is a line which have almost all zeros in the Fourier domain. This will yield ρ and effectively a and b, which gives us $H(u,v)$. We can get an estimate of $S_f(u,v)$ by computing the squared magnitude of the Fourier transform of a sub-image around the stationary object. To get an estimate of the noise variance, we can select a small homogeneous region of the image which is not blurred and compute its variance and hence the value of $S_n(u,v)$. Plugging everything in, we get $\hat{I}_{filtered}(u,v)$ and hence the filtered subimage using the inverse DFT. Seam artifacts at the subimage boundaries can be mitigated by averaging or blending.
Marking scheme: 4 points for filter in frequency domain, 4 points for commenting on spatial domain filter. 1 point for S_n, 2 points for S_f, 4 points for obtaining a, b and 1 point for final answer.

8. Let $f(x,y)$ be a real-valued 2D image of size $N \times N$ where N is a positive integer. Suppose you have an instrument that measures the Fourier transform of f at a set S of frequencies specified by the user. In other words, for $(u,v) \in S$, the instrument ideally measures $F(u,v) = \sum_{x=0}^{N}\sum_{y=0}^{N} f(x,y)\exp(-j2\pi(ux+vy)/N)$ where $j^2 = -1$, and (x,y) are discrete pixel locations. Unfortunately, due to some hardware errors, the instrument actually ends up measuring the Fourier transform at $(u+\delta, v+\delta)$ for every pair (u,v) where δ is *unknown*. Note that δ is a *single* real scalar value that is common to *all* $(u,v) \in S$.

   (a) Explain how you will estimate $f(x,y)$ and δ if the set S contains all integer-valued frequency pairs (u,v) from $-N/2$ to $N/2$.

   (b) Are there any constraints on the value of δ so that it can be uniquely estimated? Explain.

   (c) Can you uniquely determine $f(x,y)$ if it were complex-valued? Explain.

   [8+4+4=16 points]

   Solution: The instrument measures $F_\delta(u,v) = \sum_{x=0}^{N}\sum_{y=0}^{N} f(x,y)\exp(-j2\pi((u+\delta)x + (v+\delta)y)/N)$
   $= \sum_{x=0}^{N}\sum_{y=0}^{N}\big(f(x,y)\exp(-j2\pi(x+y)\delta)/N\big)\exp(-j2\pi(ux+vy)/N)$. In other words, it is measuring the Fourier transform of $g(x,y) \triangleq f(x,y)\exp(-j2\pi(x+y)\delta/N)$. Now, $g(x,y)$ can be found out using the inverse DFT and since we know that f is real, we can just ignore the phase of g to get an estimate of f.
   Due to wrapping around of phase, any value δ_1 which satisfies the following is a valid candidate for δ with k being an integer: $2\pi(x+y)\delta_1/N = 2\pi(x+y)\delta/N + k\pi$, i.e. $|\delta_1 - \delta|(x+y) = kN/2$, i.e. $|\delta_1 - \delta| = k/4$ since x

and y could both be N. Now if $|\delta_1| \leq 1/8$ and $|\delta| \leq 1/8$, we will have $|\delta_1 - \delta| \leq 1/4$ and the only way we can satisfy $|\delta_1 - \delta| = k/4$ is for $k = 0$. This shows that uniqueness holds for this constraint on δ.
If $f(x, y)$ is complex, it can be estimated only up to an arbitrary shift. This is because $\exp^{-j2\pi\delta(x+y)/N} F(u, v) = \mathcal{F}(f(x - \delta, y - \delta))(u, v)$.
Marking scheme: 8 points for the procedure to estimate f, 4 points for comments on uniqueness of δ and 4 points for stating what happens if the image is complex.

LIST OF FORMULAE:

1. Gaussian pdf in 1D centered at μ and having standard deviation σ: $p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}$.
2. 1D Fourier transform and inverse Fourier transform:
   $F(u) = \int_{-\infty}^{+\infty} f(x) e^{-j2\pi ux} dx, f(x) = \int_{-\infty}^{+\infty} F(u) e^{j2\pi ux} du$
3. 2D Fourier transform and inverse Fourier transform:
   $F(u, v) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x, y) e^{-j2\pi(ux+vy)} dx dy, f(x, y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} F(u, v) e^{j2\pi(ux+vy)} du dv$
4. Convolution theorem: $\mathcal{F}(f(x) * g(x))(u) = F(u)G(u); \mathcal{F}(f(x)g(x))(u) = F(u) * G(u)$
5. Fourier transform of $g(x - a)$ is $e^{-j2\pi ua} G(u)$. Fourier transform of $\frac{df^n(x)}{dx^n} = (j2\pi u)^n F(u)$ ($n > 0$ is an integer).