Polytechnic Institute of NYU, Dept. Electrical and Computer Engineering
EL6123 --- Video Processing, Spring 2012 (Prof. Yao Wang)
Final Exam, 5/2/2012, 3:00-5:30
One sheet of notes (double sided) allowed, No peak into neighbors. Cheating will result in failing grade!

1. (30pt) Briefly answer the following questions.
   a. In video coding, a frame is often coded as an I-frame, a P-frame, or a B-frame. Explain what does each mean. Also rank these modes in terms of coding efficiency, complexity, and encoding delay.
   b. What features of a typical video coder make the compressed bit stream very sensitive to transmission errors? List 3.
   c. What are some of the methods that can suppress error propagation after a transmission loss? List 3.
   d. A scalable video coder can offer spatial, temporal, and amplitude scalability. Explain what does each mean and how can it be achieved briefly.
   e. What are some cues that the human being uses to deduce depth of an object? List 3.
   f. In transform coding, what are the criteria for designing the transform?

2. (20 pt) Consider an image coder using block transform coding, with block size 2x2. The transform basis images $U_{k,l}$ are given below. Suppose all image samples have the same variance σ^2, and the correlation coefficient between two samples that are one pixel apart either horizontally or vertically is ρ, and the correlation coefficient between two diagonally adjacent pixels is ρ^2.

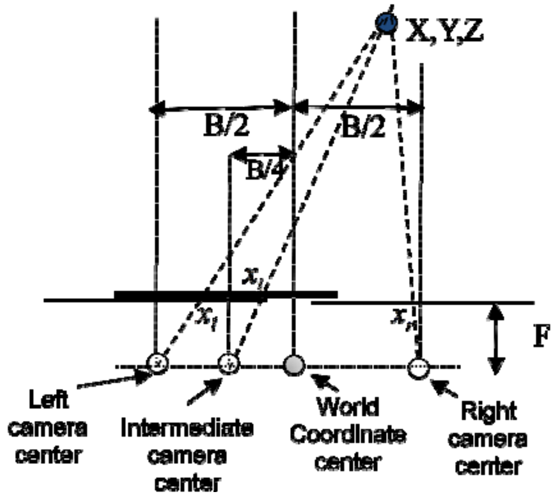
$$U_{11} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, U_{12} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}, U_{21} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}, U_{22} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix},$$

   a) Suppose we denote the transform in the form of $\mathbf{t} = [\mathbf{U}]\mathbf{s}$ where **s** is the vector containing image samples from a block, **t** is a vector containing all the coefficients, the matrix [**U**] includes the basis vectors in its columns. For this problem, you should order the samples in each 2x2 block $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ into a vector as $\mathbf{s} = \begin{bmatrix} A & B & C & B \end{bmatrix}^T$. For the given transform basis images, express the matrix [**U**] explicitly.
   b) Determine the covariance matrix of **s**.
   c) Determine the covariance matrix of ***t***.
   d) Suppose we want to use an average bit rate of *R* bits/pixel. Determine the optimal bit allocation among the coefficients. Note that you can express the solution in terms of given parameters, σ^2, ρ, R. What is the average distortion of the reconstructed image in terms of mean squares error using your bit allocation?
   e) If you have directly quantized each image sample directly with a rate of *R*. What will be the average distortion?
   f) What is the gain by using transform coding?

Note 1: you can assume the rate-distortion relation for coding either a transform coefficient or an image sample is described by $D(R) = \varepsilon^2\sigma^2 2^{-2R}$, where σ^2 is the variance of the transform coefficient or the image sample.
Note 2: For answers to (b-f), if you don't know how to find the variances of the transform coefficients, just assume they are $\sigma_k^2, k = 1,2,3,4.$ Express your solutions in terms $\sigma_k^2, k = 1,2,3,4.$

3. (20 pt) Consider a video coder using either unidirectional or bidirectional temporal prediction. Assume all the pixels have the same variance σ^2 , the corresponding pixels in two adjacent frames (frame n and frame n-1) that have a correlation coefficient of ρ and the corresponding pixels in two frames that are two-frame apart (frame n and frame n-2) have a correlation coefficient of ρ^2 . For unidirectional prediction (P-mode), a pixel in frame n is predicted from frame n-1, using a predictor of the form $f_n(x,y) = af_{n-1}(x',y')$. Here we assume that through motion estimation, we have identified that pixel (x,y) in frame n corresponds to pixel (x’,y’) in frame n-1. For bidirectional temporal prediction (B-mode), a pixel in frame n is predicted from both frame n-1 and frame n+1, using a predictor of the form $f_n(x,y) = b_1 f_{n-1}(x',y') +.b_2 f_{n+1}(x'',y'')$. Here we assume that through motion estimation, we have identified that pixel (x,y) in frame n corresponds to pixel (x’,y’) in frame n-1 and pixel (x’’,y’’) in frame n-2. (In reality, the prediction should be based on decoded values. But for ease of analysis, let us assume the prediction is based on the original values).
   a) Determine the predictor coefficient a that will minimize the prediction error in the P-mode. Also determine the prediction error variance with this predictor.
   b) Determine the predictor coefficient b_1, b_2 that will minimize the prediction error for the B-mode. Also determine the prediction error variance with this predictor.
   c) Assume the rate-distortion relation for coding the prediction error can be represented as $D(R) = \varepsilon^2 \sigma_p{}^2 2^{-\alpha R}$, where $\sigma_p{}^2$ is the prediction error variance. If we want the average distortion of the P-mode and the B-mode both equal to D_0 , what is the required bit rate (bits/pixel) for the P-mode and B-mode, respectively?
   d) If your analysis is correct, you would have found that the B-mode is more efficient. Given this analysis result, why don’t we use B-mode always in real video coders?

4. (15 pt) Consider a parallel stereo imaging system with baseline distance *B* and focus length *F* (see below). Suppose that for an object point at world coordinate *(X,Y,Z),* its image position in the left and right view are (x_l, y) and (x_r, y) , respectively.
   a. Describe how to estimate the 3D position *X,Y,Z* from x_l, x_r, y .
   b. Suppose we want to generate an intermediate view, whose camera center has a distance of *B/*4 away from the world coordinate origin, as shown below. How would you determine the image coordinate (x_i, y_i) for the same 3D point in this intermediate view? Express x_i, y_i in terms of x_l, x_r, y .
   c. Given the left and right images, briefly describe an algorithm to generate the intermediate view.



5. (15 pt) Consider the following block-based coder. Each 8x8 block in a frame is either predicted from the corresponding block in the past frame using motion compensation, or coded directly (i.e. predicted block is a constant block with values equal to 128). The coder chooses the mode that has the lowest prediction error (in terms of sum of absolute difference). Then the prediction error is coded using a transform coder, which transforms the error block using 8x8 DCT, uniformly quantizes each DCT coefficient using a specified quantization stepsize (assuming each coefficient value has a symmetric distribution around 0). The coding mode (mode=0 if predicted from the previous frame, mode=1 if coded directly), the motion vector, and the quantized DCT coefficient indices are coded using an entropy coding method. Write a Matlab code that implements the coding of a frame. Assume that the programs for motion estimation and entropy coding are given (as explained below).

   Your program should have the following syntax:

```
function [QF]=Encode(F,PF, width,height, mhmax,mvmax, QS, outfile)
```

where

F: the frame to be coded.

PF: the previous frame, previously decoded.

QF: the reconstructed frame for the current frame.

Width and Height: the width and height of a frame and assume both the width and height are dividable by 8.

mhmax,mvmax: the search range in the horizontal and vertical motions, respectively. That is, you search over a range of [-mhmax, mhmax] for the horizontal displacement, and [-mvmax, mvmax] for the vertical displacement.

Outfile: the file pointer to which to write the bitstream.

QS: the quantization stepsize for DCT coefficients

Your program can call the following functions as well as other MATLAB functions and functions defined by yourself.

```
function [mvh, mvv,PredictedBlock]=MotionEstimation(Block,RefFrame,h,v,mhmax,mvmax):
```

find the best matching block for a given 8x8 block (Block) with the top left pixel at position (h,v) in RefFrame, [mvh, mvv] are the returned motion vector components, and PredictedBlock is the best matching block. mhmax and mvmax indicate the search range for horizontal and vertical motion. Note that you don't need to write this function yourself.

```
function EntropyCoding(mode, mvh,mvv, quantDCTblock, outfile)
```

Code mode info given by "mode", the motion vector given by "mvh,mvv", and the indices of the quantized DCT coefficients of the prediction error block, given by a matrix "quantDCTblock", and write the resulting bits into a file (outfile). Note that you don't need to write this function yourself.