Module: **M1. Introduction to human and computer vision** **Final exam**
Date: November 28th, 2016 **Time: 2h30**
Teachers: Marcelo Bertalmío, David Kane, Ramon Morros, Javier Ruiz, Philippe Salembier, Verónica Vilaplana

- Books, lecture notes, calculators, phones, etc. are not allowed.
- All sheets of paper should have your name.
- **Answer each problem in a separate sheet of paper.**
- All results should be demonstrated or justified.

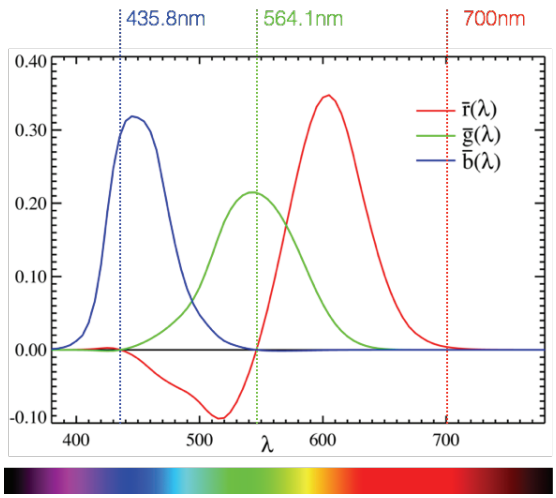
Problem I David Kane (1 point)

1. Most camera sensors can record a high bit-depth signal (>10-bit). When saving to a low bit-depth image format (typically 8-bit), the signal is first passed through an encoding nonlinearity, before the bit-depth is reduced. The image is then passed through a decoding nonlinearity when it is viewed on a monitor. Explain why the reduction in bit-depth occurs after the encoding nonlinearity and before the decoding nonlinearity.

Solution:

- The encoding nonlinearity is compressive, thus the quantization is nonlinear and finer at low luminance levels than high luminance levels.
- The human visual system is more sensitive to variations in luminance at low luminance levels.
- Thus applying quantization after the encoding nonlinearity means that the rate of quantization is well matched to that of the human visual system.
- The decoding gamma is expansive and thus returns the image to near-linearity, but with the expansive quantization remaining.
- This means that the quantization artifacts are less likely to be visible and that a lower bit depth format can be used.
- Natural images are typically low-key (they have a higher proportion of pixel in the darker half of a histogram) and compressive quantization expands the representation of low luminance values while compressing the representation of high value.

2. An additive color space describes how a limited set of base colors can be combined to produce a wider range of colors. Explain with reference to the below diagram, how the CIE color space does this and what perceptual experiments were conducted to define this space.



Solution:

- The three base wavelengths are 435.8, 564.1 and 700nm indicated by the dashed lines.

- The subject viewed with one eye light of a given wavelength, in the other eye the subject viewed the combination of the R, G and B lights.
- Their task was to manipulate the strength of R, G, and B, lights until a perceptual match was obtained.
- In the case where the wavelength matched one of the base lights (dashed lines), no other lights were needed (sanity check).
- For all other cases a combination of the R, G, and B are needed.
- However, the R gun has negative values!
- This is because for those wavelengths no combination of G and B could give rise to the same percept as the test light.
- As a result red light was added to the test light. Only then could a match be obtained.
- This problem cannot be avoided by using a different set of base wavelengths.

Problem III **Marcelo Bertalmío** **(1 point)**

1. Which visual perception property does the "automatic white balance" of cameras try to emulate? Explain the two most popular techniques for in-camera automatic white balance.
   Answer in section X of the course notes.

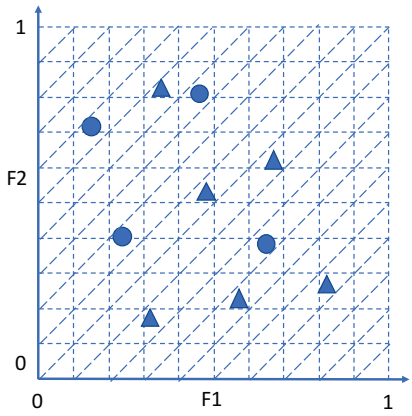
2. Why was the XYZ colorspace introduced alongside RGB? How is the function that transforms an RGB triplet into XYZ, linear or non-linear? Explain why, for any trichromatic display device, there are always colors that we can perceive but that the display is not able to reproduce.
   Answer in section III of the course notes.

Problem II **Philippe Salembier** **(2 points)**

1. In order to detect indoor images, we want to analyze the performance of two different features F1 and F2. F1 is a color-based feature and F2 a texture-based feature.
   We have an annotated database of 10 images that forms our ground truth. In involves 4 indoor and 6 outdoor images. When computing the two features on this database, we obtain the following values that are represented as points in the (F1, F2) space; circles represent indoor images and triangles outdoor images. We decide to use the following classification rule:

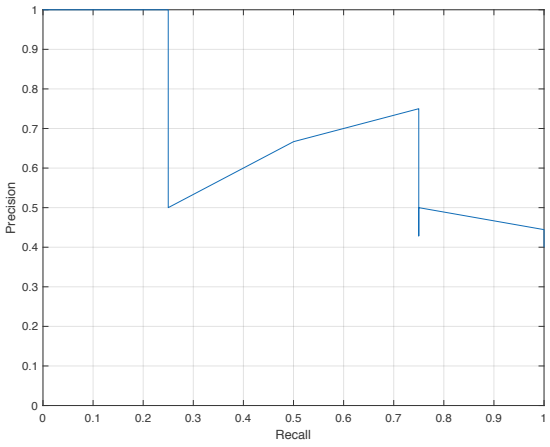
$$\begin{cases} if\ F1 - F2 \leq Th, & the\ image\ corresponds\ to\ an\ indoor\ \ scene \\ if\ F1 - F2 > Th, & the\ image\ corresponds\ to\ an\ outdoor\ scene \end{cases}$$

   Compute the Precision & Recall curve and define the optimum threshold value.



Solution:

Thresh-old	***True Positives = TP***	***Positives = P***	***True = T***	***Precision = TP/P***	***Recall = TP/T***
-0.6	*0*	*0*	*4*	*NaN*	*0/4=0*
-0.5	*1*	*1*	*4*	*1/1=1*	*1/4=0.25*
-0.4	*1*	*2*	*4*	*1/2=0.5*	*1/4=0.25*
-0.3	*2*	*3*	*4*	*2/3=0.66*	*2/4=0.5*
-0.2	*2*	*3*	*4*	*2/3=0.66*	*2/4=0.5*
-0.1	*3*	*4*	*4*	*3/4=0.75*	*3/4=0.75*
0	*3*	*5*	*4*	*3/5=0.6*	*3/4=0.75*
0.1	*3*	*6*	*4*	*3/6=0.5*	*3/4=0.75*
0.2	*3*	*7*	*4*	*3/7=0.43*	*3/4=0.75*
0.3	*4*	*8*	*4*	*4/8=0.5*	*4/4=1*
0.4	*4*	*9*	*4*	*4/9=0.44*	*4/4=1*
0.5	*4*	*9*	*4*	*4/9=0.44*	*4/4=1*
0.6	*4*	*10*	*4*	*4/10=0.40*	*4/4=1*



The best threshold is -0.1 (Precision=Recall=0.75)

2. Consider the following gray level image:
   ima =

1	1	1	1	1
1	2	2	1	3
2	2	2	2	2

2 2 2 1 2
1 2 3 1 2

Compute its erosion with the following flat structuring element $se[m,n] = \begin{bmatrix} -\infty & 0 & -\infty \\ 0 & \underline{-\infty} & -\infty \\ -\infty & -\infty & -\infty \end{bmatrix}$.

Note: The underlined element represents the space origin. If necessary assume zero padding.

Solution:

0 0 0 0 0
0 1 1 1 1
0 2 2 1 2
0 2 2 2 1
0 1 2 1 1

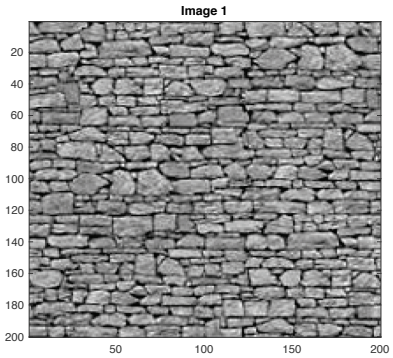
3. Define the three classical morphological gradients and describe their main difference when used for contour detection.

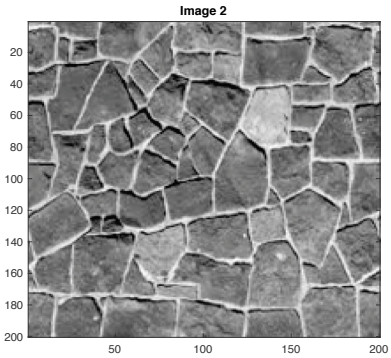
Solution:

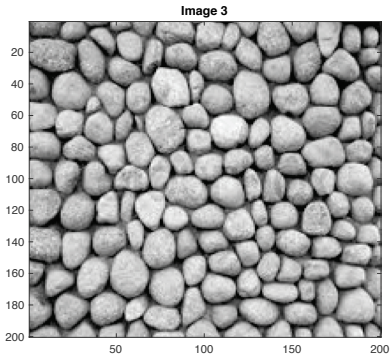
- The gradient by dilation: Dilation(f)-f. Highlight the outer contour of maxima and the inner contour of minima
- The gradient by erosion: f-Erosion(f). Highlight the inner contour of maxima and the outer contour of minima
- The morphological gradient: Dilation(f)-Erosion(f). Symmetrically highlights the contours but is 2 pixels thick.

In all cases, a small structuring element is used (3x3 or cross of 5 pixels)

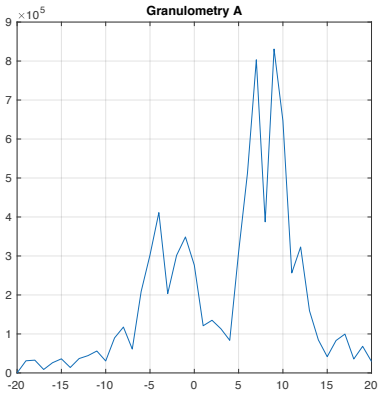
4. Considering the following three images: image 1, 2 and 3.

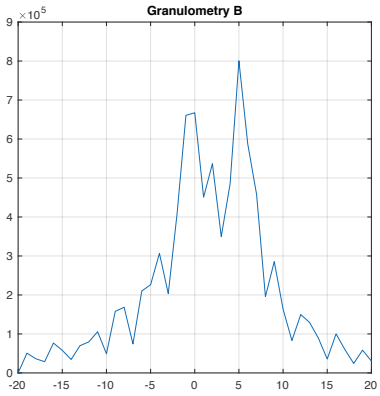


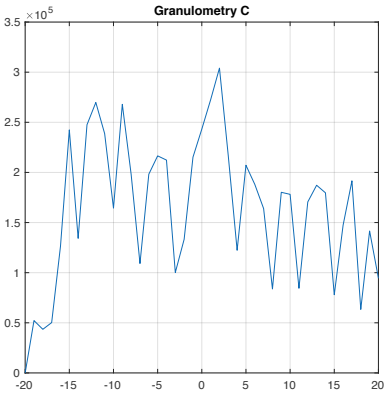




We have computed their granulometry with circular structuring element. The three pattern spectrums: granulometry A, B and C are shown below.







Define the correspondence between the granulometric curves (A,B,C) and the images (1,2,3).

Solution:

Granulometry A = image 3
Granulometry B = image 1
Granulometry C = image 2

Problem IV Javier Ruiz (3 points)

1. Using the modulation property of the DFT, express the DFT of MxN samples (M even) of the image $x[m,n] \cdot (-1)^m$ in terms of $\tilde{X}[k,l]$, the periodic version of the DFT of MxN samples of x[m,n]

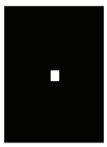
Solution: $DFT_{MxN}\{x[m,n] \cdot (-1)^m\} = x[m,n] \cdot e^{j2\pi\frac{m}{2}} = \tilde{X}[k - \frac{M}{2}, l]$

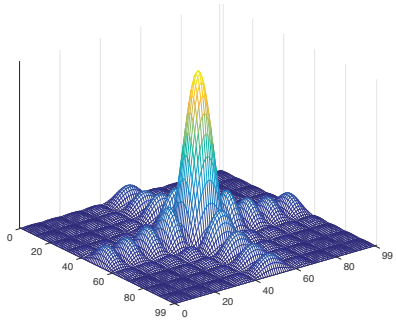
2. Compute the Fourier Transform, $X(F_x, F_y)$, of the image of MxM pixels defined by:

$x[m,n] = \delta[m] + \delta[n]$ with $\delta[k] = \begin{cases} 1 \; if \; k = 0 \\ 0 \; otherwise \end{cases}$

Solution: $FT\{x[m,n]\} = e^{-j\pi F_y(M-1)} \frac{\sin(\pi F_y M)}{\sin(\pi F_y)} + e^{-j\pi F_x(M-1)} \frac{\sin(\pi F_x M)}{\sin(\pi F_x)}$

3. Consider the image x[m,n] of 100x100 pixels composed by a black background of level 0 and a white square of level 1 of 10x10 pixels (Figure a). Figure b represents the magnitude of the Discrete Fourier Transform of 100x100 samples using the centered representation. Obtain the value of the DFT at positions X[50,50] and X[60,50].





a) Image x[m,n] b) Magnitude of the DFT with 100x100 samples

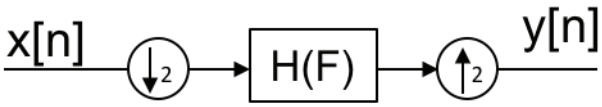
Solution: Maxim value of 10x10 =100 at k=50,l=50 (centered representation)
Zero value at k=60 and l=50 (centered) as it corresponds with the zeros of the sinc function (multiples of 1/10)

4. Detail the discrete impulse response of the Laplacian operator (second derivative). How can it be used to detect contours of an image?

Solution: $h[m,n] = \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}$ or $h[m,n] = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix}$. Detecting the zero crossings on the output.

5. Enumerate one advantage and one disadvantage of the Karhunen-Loeve Transform (KLT) versus other linear transformations.
Solution: Advantage: The KLT is a data-dependent transform which is optimal in the sense of energy compactness and therefore allows space dimensionality reduction.
Disadvantage: The basis is data-dependent and it has to be computed for each data collection.

6. Consider the following system using down-sampling and up-sampling processes and a filter with frequency response H(F). Express the Fourier Transform Y(F)=FT{y[n]} as a function of X(F)=FT{x[n]} and H(F).



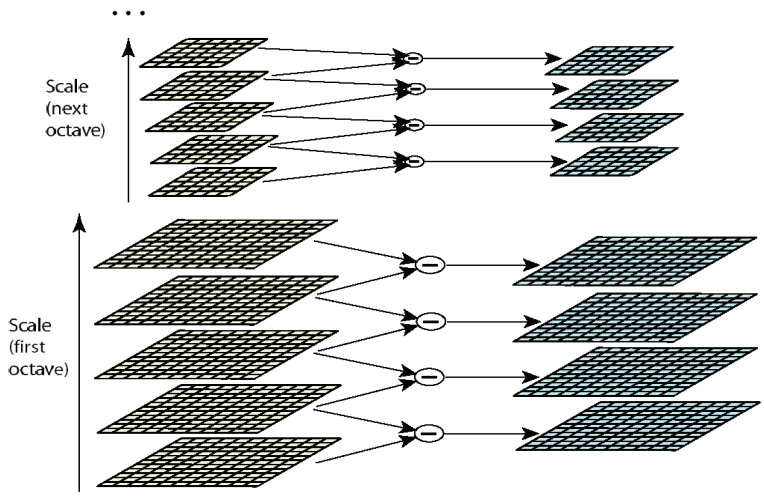
Solution: $Y(F) = {}^1/_2 X(F)H(2F) + {}^1/_2 X(F - {}^1/_2)H(2F)$

Problem V **Verónica Vilaplana** **(1 point)**

1. After running the Canny edge detector on an image, you notice that long edges are broken into short segments separated by gaps. In addition, some spurious edges appear. For each of the two thresholds (low and high) used in hysteresis thresholding, state how you would adjust the threshold (up or down) to address both problems. Assume that a setting exists for the two thresholds that produces the desired result. Explain your answer briefly.

Solution: The gaps in the long edges require a lower low threshold: parts of the long edge are detected, so the high threshold is low enough for these edges, but the edges are disconnected because the low threshold is too high. Lowering the low threshold will include more pixels of the long edges. Eliminating the spurious edges requires a higher high threshold. The high threshold should be increased only slightly, so as not to make the long edges disappear. The assumption in the problem statement ensures that this is possible.

2. The SIFT (DoG) detector finds keypoints using the image pyramid shown in the following figure. Explain how the image pyramid is created and how keypoint candidates are detected using this pyramid.



Solution:
(a) To find features that are invariant to scale, SIFT uses a Gaussian scale-space approach. Keypoints are detected by first finding scale-space extrema. This is achieved by convolving the image with Gaussian filters G at different scales of analysis σ and differencing the resulting blurred images at neighboring scales (σ and kσ for some constant k) to find local minima and maxima.
The constant k is a multiplicative factor between neighboring Gaussian-blurred images whose difference we wish to compute to extract stable features.
SIFT starts with an initial σ_0 and considers several octaves i of scale-space such that $\sigma_{i+1} = 2\sigma_i$. After each octave i, the image is downsampled by a factor of 2 in each dimension for efficiency.
More details:
SIFT subdivides each octave into s intervals, and since we want σ to double after that many intervals, it follows that $k = 2^{1/s}$, and the value of σ at octave i and interval n of the pyramid is given by $\sigma(i, n) = \sigma_0 \, 2^{i+n/s}$; $n \in [0, s-1]$
A value of s = 3 was found by Lowe to provide a good accuracy vs efficiency trade-off. The number of octaves depends on original image resolution.

(b) To extract features, SIFT compares each pixel in the DoG images to its eight neighbors at the same scale and nine corresponding neighboring pixels in each of the adjacent scales (pyramid levels). A pixel is selected as a candidate keypoint if its value is the maximum or minimum among all compared pixels.

Problem VI	**Ramón Morros**	**(2 points)**

1: Describe the LMedS algorithm and comment its advantages over RANSAC.

Solution:

- Randomly select a minimal set of points necessary to estimate an instance of the model
- The solution that produces the smallest median of residuals is picked
- Inliers scored according to their fitness to the model, outliers are given a constant weight
- The threshold is determined automatically as a value proportional to the median of the residuals produced for each subset.
- Typically produces better results than RANSAC when the inliers contain some noise because in such cases RANSAC might have thresholds set too large.

$$\min \sum_i f(r_i) \qquad f(r_i) = \begin{cases} r_i, & r_i \le t \\ t, & r_i > t \end{cases}$$

2: Describe the basic idea of segmentation on the feature space and explain all steps necessary to obtain an image partition.

Solution: Methods in this category represent the image in a selected feature space and define classes in this feature space. Pixel classification is performed according to the class assigned to its features. This defines an image classification, where each class may consist of multiple connected components. To define a partition, a labeling step in needed.

3: Formulate the problem of segmentation using GMM. Knowing that the solution can be obtained using Maximum Likelihood Estimation (Equation 1), explain the assumptions made, the parameters to compute and the procedure to obtain the solution.

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \{L(\theta|x)\} \qquad (1)$$

Solution: This is a method of segmentation of the feature space. The pdf of the feature space is modeled as a linear combination of K Gaussian density functions:

$$p(x) = \sum_{k=1}^{K} \pi_k p(x \mid \theta_k)$$

where the mixing parameters π_k can be seen as prior probabilities or normalized positive weights. The parameters to be obtained are the mixing parameters and the mean/covariance of each Gaussian distribution (μ_k, Σ_k)
Using Maximum Likelihood estimation, the parameters are computed using eq. (1), where: $L(\theta \mid x) = P(x \mid \theta)$
Solving (2) is equivalent and analytically easier.

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\{\log(L(\theta \mid x))\} \qquad (2)$$

If x = $\{x_i\}$ and assuming independence and that the density functions are Gaussians, this leads to:

$$\sum_i \log p(x_i \mid \theta) = \sum_i \log\left(\sum_k \pi_k N(x_i \mid \mu_k, \Sigma_k)\right)$$

This can be solved by taking derivatives with respect to the parameters and setting them equal to zero. When K>1 there is no closed form solution and an iterative formulation should be used.

4: Describe segmentation using a region-growing approach.

Solution: These methods define an initial, partial segmentation that do not cover all the image. The part of the image not covered is named uncertainty zone. These methods work by progressively assigning pixels from the uncertainty zone to one of the initial regions. Pixels from the uncertainty area that are connected to regions are examined and the most similar pixel is assigned iteratively. Different similarity criteria can be defined, depending on the application.
This initial segmentation is composed of a set of region 'seeds' or safe areas, that mark the interior of the final regions. These seeds are application dependent and can be obtained automatically (local maxima/minima, etc.) or manually (user interaction).