Module: **M1. Introduction to human and computer vision** **Final exam**
Date: December 2nd, 2019 **Time: 2h30**
Teachers: Ramon Morros, Javier Ruiz, Philippe Salembier, Javier Vázquez, 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 **Javier Vázquez** **(2 points)**

1. Which are the cells responding to light in the retina? Which of them are the ones focusing on color? Explain in half a page their main characteristics.

   Answer: The human retina has photoreceptor neurons, with colored pigments. These pigments have their particular photon absorption properties as a function of wavelength, and absorbed photons generate chemical reactions that produce electrical impulses, which are then processed at the retina and later at the visual cortex in the brain. The sensitivity of the pigments depends on the luminance of the light, which is a measure of the light's power, formally defined as intensity per unit area in a given direction.
   We have two types of photoreceptors:
   – Rods, for low and mid-low luminances (at high luminances they are active but saturated). There are some 120 million of them.
   – Cones, which have pigments that are 500 times less sensitive to light than the rods' pigment, rhodopsin. Therefore, cones work only with high luminances; at low luminances they are not active. There are some 6 million of them, most of them very densely concentrated at the fovea, the center of the retina. There are three types of cones: S-cones, M-cones and L-cones, where the capital letters stand for "short," "medium" and "long" wavelengths, respectively. Hubel [40] points out that three is the minimum number of types of cones that allow us not to confuse any monochromatic light with white light. People who lack one type of cone do perceive certain colors as gray.
   Frisby and Stone [38] mention that while there are several animal species with more than three types of color receptors, which can then tell apart different shades of color that we humans perceive as equal, this probably comes at the price of less visual acuity, for there are more cones to be accommodated in the same retinal area. Low luminance or scotopic vision, mediated only by rods, is therefore color-less. In a low-medium range of luminances, the so-called mesopic vision, both rods and cones are active, and this is what happens in a typical movie theatre [20]. In high-luminance or photopic vision cones are active and the rods are saturated. Each sort of cone photoreceptor has a spectral absorbance function describing its sensitivity to light as a function of wavelength: s(l);m(l); l(l). These curves were first determined experimentally by König in the late 19th century. The sensitivity curves are quite broad, almost extending over the whole visible spectrum, but they are bell-shaped and they peak at distinct wavelengths: S-cones at 420nm, M-cones at 533nm and L-cones at 584nm; see Figure 4. These three wavelength values correspond to monochromatic blue, green and red light, respectively.

2. What is colour constancy? Explain the Von Kries Law.

   Answer: Color constancy is our ability to perceive objects as having a constant color despite changes in the color of the illuminant. In 1905, Johannes von Kries formulated an explanation for color constancy that is known as von Kries' coefficient law and which is still used to this day in digital cameras to perform white balance. This law states that the neural response of each type of cone is attenuated by a gain factor that depends on the ambient light [73]. In practice, von Kries' law is applied by dividing each element of the tristimulus value by a constant depending on the scene conditions but not on the stimulus: typically, each element is divided by the corresponding element of the tristimulus value of the scene illuminant. Regardless of the original chromaticity of the illuminant, after applying the von Kries' rule the chromaticity coordinates of the illuminant become x = y = z = 1/3, which correspond to achromatic, white light. In other words, von

Kries' coefficient law is a very simple way to modify the chromaticity coordinates so that, in many situations, they correspond more closely to the perception of color.

3. List the main steps in the camera colour processing pipeline, giving a short -1 line- explanation for each of the steps.

Answer:

1. Demosaicking: An interpolation process to convert the single-channel Bayer-patterned image into a full color image.
2. Black level adjustment: Discount the "black level", that is caused by Dark Current, Dark current is a parasite current, not originated from photon conversion, that nevertheless is integrated as charge. To do so, the borders of image arrays are composed of "optical black pixels," which are never exposed so that they can be used to estimate dark current levels and therefore a proper black level for the image.
3. White balance: This step tries to mimic our color constanct ability, i.e. our ability to perceive the color objects stable under different lighting conditions.
4. Color correction: This step converts the RGB values of the camera into the standard CIE XYZ color values. In this way, this step links our captured image with the human visual perception.
5. Color correction (2): This step converts the CIE XYZ values to the output RGB color space.
6. Gamma correction: This step introduces a non-linearity that serves 2 purposes: First, it renders the image so it can be displayed in the monitor, and also, it does so by preserving better those regions to which we are more sensible.
7. Noise reduction: This step aims at discounting all the noise introduced by the sensor, and that is propagated through the camera pipeline.
8. Contrast and color enhancement: This step focus on obtaining more pleasing images, using things like edge enhancing.
9. Compression: This is the last step. The image is prepared for storage and broadcast by reducing it to just 8 bytes per pixel. This reduction is done through quantization.

4. Explain unsharp masking. Why is it needed in the camera processing pipeline?

Answer: Unsharp masking is a basic edge enhancement method, i.e. it aims at enhancing the edges present in the image to give a "sharper" look. Unsharp masking is linear, and it works in the following way.
From an image I, we compute an edge map E, by the difference of the original image and the result of convolving the original image with a gaussian E = I - g * I. Then, this edge map is added to the original image with a scaling factor k
I'=I+K·E

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

1. In this problem, our goal is to evaluate the performances of two image classification systems, which have been designed to differentiate between urban images and non-urban images. In order to perform an objective assessment a small database of images has been created. It is composed of the following 10 images.


Image 1


Image 2


Image 3

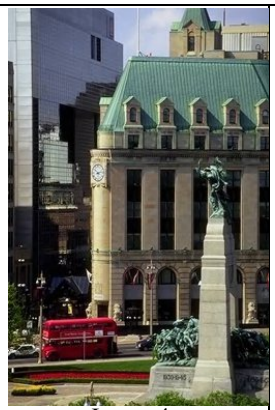


Image 4


Image 5


Image 6


Image 7

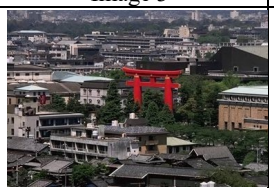


Image 8


Image 9


Image 10

It has been decided to assess the performances of the systems with the F measure which is the geometric mean of Precision, P, and Recall, R; that is $F = 2.P.R/(P + R)$

The two systems are run on the database and give the following classification results:

System	Image 1	Image 2	Image 3	Image 4	Image 5	Image 6	Image 7	Image 8	Image 9	Image 10
A	Urban	Non-urb	Urban	Urban	Urban	Non-urb	Urban	Urban	Non-urb	Urban
B	Non-urb	Non-urb	Urban	Urban	Urban	Non-urb	Non-urb	Urban	Non-urb	Urban

Note that both systems make two incorrect classifications. Both classify *Image 5* as an urban image, which is not the case. Moreover, System A is based on color information and misclassifies *Image 7* assuming that we see a lot of areas made of concrete. On its turn, System B, which is based on texture, estimates that *Image 1* is from a non-urban scene.

Assuming that the class of **"True" samples is the class of Non-urban images**, compute the precision, recall and F values of both systems and decide which one has the best performances.

Solution:
Ground truth: True=5, False=5
System A: Positive=3, Negative=7, True Positive=3, Precision=1.0, Recall=0.6, F=0.75
System B: Positive=5, Negative=5, True Positive=4, Precision=0.8, Recall=0.8, F=0.80
System B has better performances that system A.

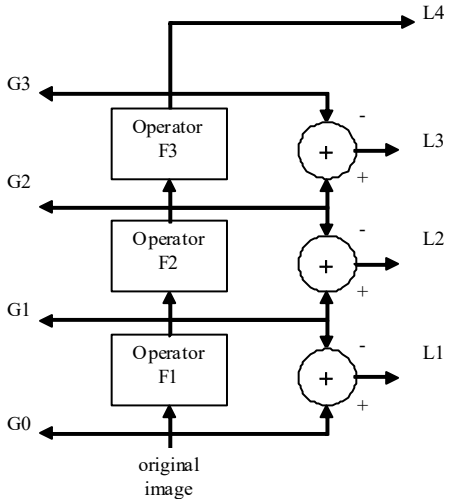
2. Consider the following range transform: $s = T(r) = 4(\mathrm{r} - \frac{1}{2})^2$, for $0 \leq r < 1$
   Justify whether this transform increases, maintains or reduces the entropy of discrete images.

Solution:
The transform reduces the entropy because two different gray levels at the entrance have the same gray level at the exit

3. State the three important algebraic properties of a closing and define mathematically each property.

Solution:
Increasing: if $x \leq y \implies \gamma(x) \leq \gamma(y)$
Extensive: $\forall x, x \leq \gamma(x)$
Idempotent: $\forall x, \gamma(\gamma(x)) = \gamma(x)$

4. Multiscale image decomposition can be performed with the scheme of the following figure. In the sequel, various schemes are created with several operators Fi. In all cases, you can ignore the image border effects (that is you can consider that the image is of infinite size).



a) Demonstrate that, for any operators Fi, the sum of the Li images is always equal to the original image.
b) Assume that F1 is an opening, F2 a closing and F3 a closing of opening (closing(opening(.)). All operators use the same flat structuring element of size 5x5. Describe the content of the Li images.

c) Assume that F1 is an erosion with a horizontal flat structuring element 1x3, F2 an erosion with a vertical flat structuring element 3x1 and F3 a dilation with a square structuring element of size 3x3. Is there any order relationship between L4 and G0?

d) Assume that F1 is an opening with a flat structuring element of size 5x5, F2 an opening with a flat structuring element of size 3x3 and F3 a closing with a flat structuring element of size 5x5. Describe the content of the Li images.

Solution:
a) L1= G0-G1, L2 = G1-G2, L3 = G2-G3, L4 = G3.

So: L1+L2+L3+L4 = (G0-G1) + (G1-G2) + (G2-G3) + G3 = G0

b) L1 contains the bright image elements of size lower than 5x5.
L2 contains the dark image elements of size lower than 5x5.
L3=0 (Close(open(.)) is a morphological filter, therefore idempotent).
L4 contains the image elements of size larger than 5x5.

c) The concatenation of the first two erosions is an erosion with a structuring element of size 3x3. Therefore, L4 can be seen as the result of an opening, which is anti-extensive; therefore
L4 is lower than or equal to G0.

d) L1 contains the bright image elements of size lower than 5x5.
L2=0 (see chapter on granulometry: the combination of two opening of different sizes but with convex structuring elements is equal to the one of largest size).
L3 contains the dark image elements of size lower than 5x5.
L4 contains the image elements of size larger than 5x5.

Problem III **Javier Ruiz** **(3 points)**

1. Consider the following LSI filter with impulse response (kernel) of size 3x3: $h[m,n] = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix}$ and the image $x[m,n] = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$ of size 4x4.

   a. What is the size of the filtered image $y[m,n] = x[m,n] * h[m,n]$?
   b. Compute the values of the filtered image $y[m,n] = x[m,n] * h[m,n]$ (If necessary zero-padding may be assumed).
   c. Does the filter correspond to a low-pass or high-pass filter? In which (horizontal or vertical) component?
   d. Justify if the filter detects any vertical or horizontal contours.

Solution:

a. 6x6

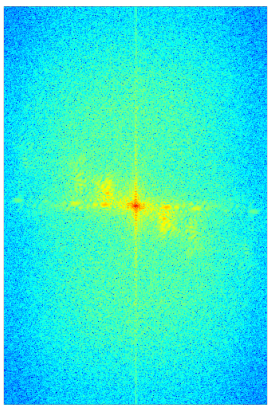
b.
0	0	1	1	-1	-1
0	0	2	2	-2	-2
0	0	3	3	-3	-3
0	0	3	3	-3	-3
0	0	2	2	-2	-2
0	0	1	1	-1	-1

c. The filter corresponds to a high pass filter on the horizontal component

d. vertical contours relate to high frequency in the horizontal component, therefore they will be detected

2. We want to analyse the frequency content of an image of 330x500 pixels (Figure a) with its DFT-2D of 330x500 samples (Figure b):

(a) Original image

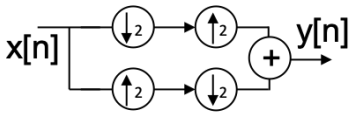


(b) DFT-2D magnitude

   a. What discrete frequencies (F1 horizontal and F2 vertical) correspond to the DFT sample located at column 202 / row 251? column 130 / row 251? column 202 / row 271 and column 130 / row 231? (remember that Figure b shows the centered representation of the DFT)
   b. What region of the image corresponds to those frequencies?

Solution: a. c=202,r=251 → F1=0.1,F2=0
c=130,r=251 → F1=-0.1,F2=0
c=202,r=271 → F1=0.1,F2=0.04
c=130,r=231 → F1=-0.1,F2=-0.04
b. The first two samples correspond to the left part of the girl blouse (as vertical F2 is 0). The other two correspond to the right part of the blouse (diagonal frequencies as F1 and F2 are different than 0).

3. Consider the following decomposition using down-sampling and up-sampling processes (without filtering). Express the Fourier Transform Y(F)=FT{y[n]} as a function of X(F)=FT{x[n]}.



Solution: $Y(F) = {}^{3}/_{2}\,X(F) + {}^{1}/_{2}\,X(F - {}^{1}/_{2})$

4. Given a decomposition of an image X into a Laplacian pyramid elements with 3 levels: L_1, L_2 and G_3.

a. Comment briefly the differences between Gaussian and Laplacian pyramids.
b. Compute the number of samples (pixels) of the Laplacian representation for an image of size NxN.
c. Is the Laplacian pyramid a complete or overcomplete representation of the image X?

Solution:
a. The Gaussian pyramid consists of a series of images that are iteratively filtered using a Gaussian filter and scaled down. The Laplacian pyramid is similar to the Gaussian but it computes the difference between up-sampled Gaussian pyramid level and the Gaussian pyramid level effectively. It represents a band pass filter (except for the last level).
b. NxN + N/2xN/2 + N/4xN/4
c. Overcomplete as it is greater than NxN

5. Discuss the advantages and disadvantages of the Discrete Cosine Transform (DCT) versus the Karhunen-Loeve Transform (KLT).

Solution: The DCT is a complete, separable and orthogonal transform whose transformed coefficients are real and present good compaction characteristics (better than DFT).
The KLT is a data-dependent transform which is optimal in the sense of energy compactness and therefore allows space dimensionality reduction. The basis is data-dependent and it has to be computed for each data collection

6. Enumerate two main ways to reduce the resolution of images/features at each layer of convolutional neural networks?

Solution: strides when filtering (convolution) and adding pooling layers

Problem IV Verónica Vilaplana (1 point)

1. Compare and contrast Harris and Difference of Gaussians (DoG) detectors.

Answer:
Harris is a corner detector, DoG is a 'blob' detector.
The two detectors are covariant to rotation, translation and (partially) invariant to intensity changes. Harris is not invariant to scaling while DoG is scale invariant.
Harris blurs the image first with a Gaussian filter to remove noise, then computes a cornerness function that depends on the second moment matrix, applies a threshold and non-max suppression; DoG builds a difference of Gaussians pyramid, convolving the image with Gaussian filters of increasing width, and computing differences of consecutive filtered images. Then it if finds points with max response, both in scale and space.

2. Assuming feature points have been previously detected using the SIFT feature detector, briefly describe the main steps of creating the SIFT feature descriptor at a given feature point.

Answer:
At each point where a SIFT "keypoint" is detected, the descriptor is constructed by computing a set of 16 orientation histograms based on 4 x 4 windows within a 16 x 16 pixels neighborhood centered around the keypoint. At each pixel in the neighborhood, the gradient direction (quantized to 8 directions) is computed using a Gaussian with σ equal to 0.5 times

the scale of the keypoint. The orientation histograms are computed relative to the orientation at the keypoint, with values weighted by the gradient magnitude of each pixel in the window. This results in a vector of 128 (= 16 x 8) feature values in the SIFT descriptor. (The values in the vector are also normalized to enhance invariance to illumination changes.)

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

1. Describe how the gradient can be used to improve the Hough algorithm for detecting shapes in images. Explain the benefits of using the gradient in Hough transform.
   Answers:
   For each edge, the gradient phase already provides the orientation of the edge. There is no need to loop over all possible angles.
   This makes the process much faster.

2. In the RANSAC algorithm, what is the relationship between the number of outliers and the necessary number of iterations?
   Answers: The number of necessary iterations increases with the number of outliers

3. Schematically explain the Max-Lloyd algorithm used to perform the k-means clustering.
   Answers:
   a Initialize K classes. Compute the centers of each class
   b For each point:
   - Compute the distances between the point and the class centers
   - Assign the point to the closest class
   c Update the class centers
   d Repeat b) & c) until no change (in assignments or center values) is observed.

4. Explain briefly the difference between region merging and region growing methods in segmentation.
   Answers:
   Region merging: we start from a partition (all pixels belong to a connected region with unique label) of the image. Neighboring regions are iteratively merged using a region similarity criterion until a given stop criterion is reached.
   Region growing: we starting from a partial segmentation, defined by some markers (‘safe’ areas defining the interior of the final regions) and an uncertainty zone (unlabeled pixels, not belonging to any region). The pixels in the uncertainty zone connected to the markers are iteratively assigned to the regions using a similarity criterion, until all pixels in the uncertainty zone are labeled (assigned to a region).