Module: **M4. Video analysis** **Final exam**
Date: February 19th
Teachers: Montse Pardàs, Ramon Morros, David Varas, Constantine Butakoff, Ferran Marqués, Javier Ruiz, Josep Ramon Casas, Jordi González **Time: 2h**

- 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 1: **2 Points**

In order to decompose a video sequence into various shots using a transition based segmentation approach, different features can be computed for each frame: the histogram, the Frame Difference $FD = \sum_{\vec{r}} \left(I(\vec{r},t) - I(\vec{r},t-\Delta t) \right)$ or the Displaced Frame Difference $DFD = \sum_{\vec{r}} DFD\left(\vec{r}, \hat{D}(\vec{r})\right) = \sum_{\vec{r}} \left(I(\vec{r},t) - I(\vec{r} - \hat{D}(\vec{r}),t-\Delta t) \right)$.

1. Describe the algorithms for producing the shot segmentation using each one of these features.
2. Indicate which feature may produce the largest amount of false shot transitions in case of strong motion.
3. Indicate which feature is the most appropriate to detect shot transitions that are produced with gradual transitions using geometric effects, like in the following example:

4. Propose a modification of one of the three features mentioned above to make the algorithm robust to smooth illumination changes within a shot.

1. The FD, DFD and a measure of the distance between histograms of successive frames would be computed for each frame. This would produce a one dimensional function where we could apply a threshold to detect shots.
2. FD
3. FD, or histogram based, using a low threshold.
4. We could use a histogram of the chrominance values, or modify the DFD estimation in order to estimate the illumination changes as well.

Problem 2: **2 Points**

Assume you have an outdoor video sequence recorded with a still camera, that you started to capture when there was no person in the scene. The background is dynamic due to the wind, and there are illumination changes typical of outdoor scenes. This is one frame of the sequence:

Describe two algorithms that you can use to extract the foreground objects. These algorithms should be able to model the dynamic background and extract only the person.
Background modeling with a Gaussian Mixture Model, Kernel density estimation, or eigenbackground. See the slides for the description of these algorithms.

Problem 3: **1 Point**
When using block-matching, explain the differences between forward and backward motion estimation
See slides of the course, pp. 29-32

Problem 4: **1 Point**
The motion vectors obtained using the Block Matching algorithm may not accurately represent the motion of the objects in the scene. Explain two reasons that may cause failure at obtaining the true motion.
1) One block may have multiple possible references in the previous image (e.g. uniform regions). 2) Assumption that all the pixels in a block move with a translational motion is not true.

Problem 5: **1 Point**
Given a set of eigenvalues (6,5,4,3,2,1) that correspond to 6 principal components of the PCA, calculate the variability (as percentage of total variability) explained by the first two principal components (you can leave the answer as a fraction).
Total variability is measured by the sum of all the eigenvalues. Total variability = 6+5+4+3+2+1=21
Variability explained by the first 2 principal components is 100*(6+5)/21=100*11/21

Problem 6: **1 Point**
Describe the steps of the Procrustes Analysis (alignment) and what its purpose is in the framework of active shape model.
Procrustes analysis is used to align the shapes to remove from the dataset variability due to a similarity or affine transformation. It is required to avoid modeling the shape deformation due to the transformation (e.g. rotation, translation).
Steps:
1. Define starting average shape (could be one from the set of sshapes)
2. Center it at (0,0), rescale to unit size (such that |x|=1 with x being the shape)
3. Align all the shapes to the average shape
4. Calculate the average of the aligned shapes
5. Normalize the average to unit size
6. Repeat from step 3 until the average does not change any more

Problem 7: **1 Point**
In a Particle Filter framework, explain what the degeneracy phenomenon is. Is it possible to avoid its appearance? What is the step of the filter that minimizes the effect of this phenomenon?
The degeneracy phenomenon is the effect caused by the concentration of most of the probability mass in one particle. It is not possible to avoid, but it can be minimized using a resampling step.

Problem 8: **1 Point**
Consider a Kalman filter used to estimate the horizontal position of a car moving with constant velocity (v).

a) Assuming that the state vector ($\vec{x}_t$) is composed by the position and the velocity of the object at each time instant, the system dynamics of the car can be expressed as:

$$\vec{x}_t = \boldsymbol{D} \cdot \vec{x}_{t-1} + \vec{n}$$

Comment whether the system dynamics follow a Linear Dynamic Model (LDM) or not depending on the previous expression and the noise vector.
The expression shows a linear transformation plus the addition of noise. Thus, if the noise is Gaussian, the dynamics of the system follow a LDM.

b) Is the Kalman Filter a good estimator for this problem? How many parameters should be estimated to characterize the dynamics of the system using a Kalman Filter?
If the noise is Gaussian, the Kalman filter is optimal for this problem. In a Kalman filter framework, 2 parameters should be estimated at each iteration, mean and variance of a Gaussian function.

c) Explain the steps followed by the Kalman Filter to estimate these parameters. Comment the behaviour of the system when the uncertainty of the measurement or of the prediction is negligible.
The two steps followed by the Kalman Filter are Prediction (estimation without measurement) and Update (once the measurement is available).
Uncertainty of the measurement is negligible -> only the measurement is used
Uncertainty of the prediction is negligible -> only the prediction is used

Problem 9: **2 Points**

In sport events, the knowledge of the ball position carries valuable information about the game. In particular, in football matches, the position and trajectory of the ball is crucial to analyze the game. In this exercise, ball tracking in a football match is tackled. Two different game situations are considered in this analysis.

Figure 1

Figure 2

In the first situation (Figure 1), the ball is represented by a set of image pixels

a) Suppose that we decide to represent the ball using a circle. Write the expressions of the state vector in this particular case and the estimation of its pdf using a generic particle filter.
A Particle Filter is a train of deltas multiplied by their associated weights:

$$p(x_k \mid z_k) \approx \sum_{i=1}^{N} w_k^i \cdot \delta\left(x_k - x_k^i\right)$$

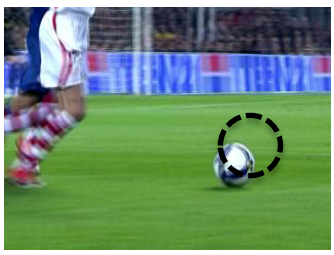
In this case, each estimation of the ball can be parametrized by the center of a circle and its radius.

b) Explain the relation between measurement and states during the tracking process.
The states represent the real position of the ball and are hidden during the tracking process. Measurements are the information to which we have access and are related with the states. We use the information in the estimates represented by circles to estimate the real states.

c) Consider three particles (dashed) tracking the ball presented in Figure 1:

Particle 1

Particle 2



Particle 3

Knowing that the normalized weights of these particles are {0.9, 0.09, 0.01}, associate each weight with its particle. Compute the Effective Sample Size, comment its maximum/minimum and when they are achieved. Discuss whether it is necessary or not to introduce a resampling step in this iteration.

In the second situation (Figure 2), consider the ball occluded by a player legs.
Particle1-> 0.01
Particle2-> 0.9
Particle3-> 0.09

ESS = 1/((0.01)^2 + (0.9)^2 + (0.09)^2) ≈ 1/0.81 = 1.2 < 1.5
From the expression above we can conclude that a resampling step is needed.

Min ESS = 0; Max ESS = N

d) Explain the difference between the Kalman Filter and the Particle Filter in terms of number of estimations handled at the same time. Can multiple estimations be usefull when the ball appears again?
Kalman Filter generates a single estimation at each time instant, whereas the Particle Filter handles multiple estimations associated with the particles. It can be useful to provide multiple estimations when the ball appears again because we don't know its movement and its trajectory while it was occluded.

Problem 10: **1 Point**

In the compression context, explain which features are desirable in a transform. Discuss which of such features are ensured by the K-L Transform and which by the Discrete Cosine Transform (DCT) and, in both cases, how these features happen to be ensured.

Desirable features for a transform are in Slide 25.
KLT properties are in Slide 31 and Slide 32.
DCT properties are in Slide 38 and Slide 41-42.

Problem 11: **1 Point**

In the case of a B-frame in the standard MPEG, explain the three different types of prediction that can be used and discuss the usefulness of each one. How many motion vector solutions are computed in an exhaustive search for a MacroBlock in a B-frame? Assume that the MacroBlock has size 16x16 pixels and the maximum allowed displacement in any direction is of P = 8 pixels.

The three different modes are in Slide 102.
The mechanism to compute the number of motion vectors is in Slide 86.

Problem 11b: **1 Point**

Explain briefly the main blocks of a gesture recognition system.
A: lesson 8, slide 10: Capture, Gesture Analysis (Segmentation, Feature extraction, Training, Model & Feature prediction) and classification.

Problem 12: **1 Point**

Discuss the difficulties of marker-less based motion capture systems.
A: lesson 8, slide 23: Poor imaging: motion blurred, occlusions, bad correspondences and multimodal mapping

Problem 13: Model-based tracking **2 Point**

1) What are the main tasks in the framework of human motion analysis (HMA)?
2) What is *Mocap* and how is it used in media production?
3) Is it possible to detect/recognize a small set of actions without using an explicit human body model in the detector?
4) Model-based tracking allows for dimensionality reduction, how?
5) But, even with the dimensionality reduction imposed by the body model, exhaustive exploration of the search space is impractical in the estimator. Name a few solutions to overcome this problem.

1) Motion capture (estimation and tracking), action and behavior recognition, segmentation of human motion
2) Motion capture (mocap) consists in recording human movement and translating it onto a digital model. It is used for the animation of avatars with real human motion parameters of the performers
3) Yes. For example extracting low level features (such as salient points positions or optical flow) and training a classifier with these trajectories/flows in the apparent image
4) Because the search in the estimator can be reduced to the space of parameters defining the human body model (HMB). E.g. all the possible configurations of positions and angles of the joints, limited by kinematic constraints of the articulated HBM
5) Annealed Particle Filtering (Deutcher 2005), Partitioned Sampling (MacCormick 2000), Hierarchical Particle Filtering (Bandouch 2009), Layered Particle Filtering (López 2012)

Problem 14: **2 Point**

Briefly describe and compare the top-down vs. bottom-up behavior models. For each type of methodology, please detail the advantages/drawbacks of each method, and illustrate in which type of behavior analysis applications are best suited.
Bottom-up: a behavior model is built by analyzing a set of observations.

Advantages/drawbacks of bottom-up:
- A set of normal behavior patterns is learnt
- The model can change over time if new patterns are observed
- More robust to noise (if considered in the learning step)
- No semantic explanation can be extracted
- Difficult to interpret by users

Applications of bottom-up:
- virtual tripwires (illegal turns)

- moving in abnormal directions (flow control)
- human body actions

Top-Down: the behavior model is specified beforehand.

Advantages/drawbacks of top-down:
- All the knowledge must be provided by an expert
- Not robust to noisy observations
- Useful for restricted environments. Not evolvable
- Can provide accurate semantic descriptions
- Easy to understand by users

Current applications:
- object left behind (abandoned objects)
- object inside (subway)
- crowded density
- person counting
- vandalism